Mangalore University

Exploring Predictive Models for Analysis of the Epidemic Status of HIV/AIDS

Pre-Synopsis

of the PhD Thesis to be submitted for the award of degree of

Doctor of Philosophy

in

COMPUTER SCIENCE

By:

Daniel Mesafint Belete
Research Scholar

Under the Guidance of:

Dr. Manjaiah Doddaghatta Huchaiah
Professor

Department of Post-Graduate Studies and Research in Computer Science,
Mangalore University,
Mangalagangotri - 574199

October, 2021

Contents

List of Figures iii

List of Tables v

1 Introduction and Literature Survey 1
- 1.1 Introduction . . . 1
  - 1.1.1 Overview of HIV/AIDS . . . 1
  - 1.1.2 Status of HIV Epidemic in Ethiopia . . . 1
  - 1.1.3 Ethiopia Demographic and Health Survey (EDHS) . . . 3
  - 1.1.4 Literature Review . . . 3
- 1.2 Motivation . . . 10
- 1.3 Objectives of the Research . . . 10
- 1.4 Justification of the Research . . . 11

2 Research Methodology 12
- 2.1 Dataset Collection and Pre-processing . . . 12
  - 2.1.1 Data Source . . . 12
  - 2.1.2 Data Preprocessing . . . 13
    - 2.1.2.1 Data Integration . . . 13
    - 2.1.2.2 Data Cleaning . . . 13
    - 2.1.2.3 Handling Missing Value . . . 13
  - 2.1.3 Data Transformation . . . 14
    - 2.1.3.1 Data Discretization . . . 14
    - 2.1.3.2 Data Reduction . . . 14
    - 2.1.3.3 Data Normalization . . . 14

3 Feature Selection Methods for Prediction of the Individual's Status of HIV/AIDS from EDHS Dataset- A Filter Approach 16
- 3.1 Introduction . . . 16
- 3.2 Filter Based Feature Selection Methods . . . 16
  - 3.2.1 Univariate Feature Selection . . . 16
  - 3.2.2 Feature Importance . . . 16
  - 3.2.3 Correlation Coefficient . . . 17
- 3.3 Experimental Results . . . 17
- 3.4 Conclusion . . . 20

4 Wrapper Based Feature Selection Techniques on EDHS-HIV/AIDS Dataset 21
- 4.1 Introduction . . . 21
  - 4.1.1 Forward Feature Selection (FFS): . . . 21
  - 4.1.2 Backward Feature Selection (BFS): . . . 21
  - 4.1.3 Recursive Feature Selection (RFS): . . . 21
- 4.2 Experimental Results . . . 22
- 4.3 Conclusion . . . 24

5 A Comparative Study of Filter and Wrapper Methods on EDHS – HIV/AIDS Dataset **25**
5.1 Introduction . . . 25
5.2 Proposed Methodology . . . 25
5.2.1 Preprocessing . . . 25
5.3 Experimental Results and Performance Evaluation . . . 26
5.3.1 Results from full feature set . . . 26
5.3.2 Results from Features Selection . . . 27
5.3.2.1 Wrapper . . . 27
5.3.2.2 Filter . . . 28
5.3.3 Performance Evaluation . . . 28
5.4 Conclusion . . . 29

6 Performance Evaluation of Classification Models for HIV/AIDS Dataset **30**
6.1 Introduction . . . 30
6.2 Proposed Methods . . . 30
6.3 Experimental Result . . . 30
6.3.1 Comparison with previous work . . . 32
6.4 Conclusion . . . 33

7 Classification of Binary Class HIV/AIDS Test Results Using Ensemble Learning Models **35**
7.1 Introduction . . . 35
7.2 Proposed Methodology . . . 35
7.3 Experimental Results and Discussion . . . 35
7.4 Conclusion . . . 39

8 Grid Search In Hyperparameter Optimization of Machine Learning Models for Prediction of HIV/AIDS Test Results **40**
8.1 Introduction . . . 40
8.2 Proposed Methodology . . . 40
8.3 Experiments . . . 41
8.3.1 Experimental Setup . . . 41
8.4 Results and Discussion . . . 42
8.5 Conclusion . . . 46

9 Deep Learning Models for Predicting HIV Test Status using DHS data **49**
9.1 Introduction . . . 49
9.2 Proposed Methodology . . . 49
9.3 Experimental setup and result discussion . . . 49
9.3.1 Experimental Setting . . . 50
9.3.2 Experimental Result Analysis . . . 50
9.4 Comparison . . . 53
9.5 Conclusion . . . 55

List of Publications **56**

List of Figures

1.1 The HIV virus structure [1] . . . 1
1.2 HIV Prevalence among adult men and women by Region in Ethiopia EDHS 2016 [2] . . . 2
1.3 Proportion of People Living with HIV (PLHIV) by Region in Ethiopia [2] . . . 2

3.1 Selected features using feature importance criteria . . . 17
3.2 Heatmap of Pearson correlation coefficient matrix for the considered EDHS dataset . . . 18

5.1 The proposed methodology . . . 25
5.2 The ROC curve on full training set (feature set) . . . 27
5.3 The ROC curve on wrapper methods . . . 27
5.4 ROC curve on filter methods . . . 28

6.1 Proposed Methodology . . . 30
6.2 Comparison of ten classifier with AUC/ROC. Area under curve (left), ROC curve (right) . . . 31
6.3 Comparison of ten classifier with Precision and recall. Precision (left), Recall (right) . . . 32
6.4 Results of each classifier F1-score. . . . 32
6.5 Confusion matrix the HIV/AIDS dataset. . . . 33

7.1 Proposed methodology of ensemble learning method . . . 36
7.2 ROC/AUC results of the five models before and after feature selection: before (left) , after (left) . 37
7.3 Precision and recall results of the five models before and after feature selection: before (left), after (right) . . . 37
7.4 The learning curve of Gradient Boosting (GB) before and after feature selection: before (left), after (right). . . . 37
7.5 The learning curve of Multilayer Perceptron (MLP) before and after feature selection: before (left), after (right). . . . 38
7.6 The learning curve of Random Forest (RF) before and after feature selection: before (left), after (right). . . . 38
7.7 The learning curve of Extra Tree (ET) before and after feature selection: before (left), after (right). 38
7.8 The learning curve of K-nearest neighbor (KNN) before and after feature selection: before (left), after (right). . . . 38
7.9 Confusion Matrix for five established classifiers after feature selection applied. . . . 39

8.1 Flow diagram of the proposed grid search hyper-parameter optimization (GSHPO) method . . . 41
8.2 The feature importance of Random Forest and Decision Tree : RF (left), DT (right) . . . 45
8.3 The feature importance of Extra Tree and Gradient Boosting : ET (left), GB (right) . . . 45
8.4 The feature importance of AdaBoost . . . 45
8.5 The comparison of each model with ROC Curve and Precision - Recall Curve: ROC (left), Precision - Recall (right) . . . 46
8.6 Confusion matrix results based on training data of the established mode . . . 47
8.7 Bias-variance tradeoff of the model with learning curve . . . 48

9.1 The proposed deep learning approach's architecture. . . . 50
9.2 Deep learning Models AUC using the train-test split strategy: ANN (left), CNN (right) . . . 52
9.3 Deep learning Models AUC using the train-test split strategy: RNN (left), LSTM (right) . . . 52
9.4 Deep learning Models AUC using the train-test split strategy: CNNLSTM (left), CNNRNN (right) . . . 52

9.5 Prediction accuracy on the proposed deep learning model on HIV/AIDS datasets: ANN (left), CNN (right) . . . 53
9.6 Prediction accuracy on the proposed deep learning model on HIV/AIDS datasets: RNN (left), LSTM (right) . . . 53
9.7 Prediction accuracy on the proposed deep learning model on HIV/AIDS datasets: CNNLSTM (left), CNNRNN (right) . . . 53
9.8 The loss caused by the proposed technique on deep learning models on HIV/AIDS datasets: ANN (left), CNN (right) . . . 54
9.9 The loss caused by the proposed technique on deep learning models on HIV/AIDS datasets: RNN (left), LSTM (right) . . . 54
9.10 The loss caused by the proposed technique on deep learning models on HIV/AIDS datasets: CNNLSTM (left), CNNRNN (right) . . . 54

List of Tables

1.1 Recent papers from standard publications are surveyed, reviewed and summarized in the table below. . . . 4

2.1 The Total records of the EDHS dataset of four separate year of Survey . . . 12
2.2 The Initial records of the HIV/AIDS dataset of Four Separate year from EDHS Survey . . . 13
2.3 Encoding new value for replacement of existing value . . . 15

3.1 Selected features using univariate feature selection methods . . . 17
3.2 Result of filter based feature selection methods on EDHS-HIV/AIDS dataset . . . 18

4.1 Result of Wrapper based feature selection methods on EDHS-HIV/AIDS dataset . . . 22

5.1 Characteristics of the EDHS-HIV/AIDS Dataset . . . 26
5.2 Results of both classifier on full training set with original feature set . . . 27
5.3 Result of wrapper feature selection on both classifiers. . . . 28
5.4 Result of filter feature selection on both classifiers. . . . 28
5.5 Performance result of full data set, filter set and wrapper set on both classifiers . . . 29

6.1 Results of Classification algorithm based on different performance metrics . . . 31
6.2 Comparison of the proposed method with selected related research works with respect to accuracy 34

7.1 Classification report of all models before and after feature selection . . . 36

8.1 The 10-fold cross validation score for all models before optimized . . . 41
8.2 The machine learning models hyperparameter configuration space . . . 42
8.3 The 10-fold cross validation score of all models after applying Hyperparameter optimization . . . 43
8.4 Results of evaluation according to the standard metrics for each model before Grid Search technique is applying . . . 44
8.5 GSHPO results of evaluation according to the standard metrics for each model . . . 44

9.1 The outcomes of all DL models evaluated using a 10-fold cross-validation methodology. . . . 50
9.2 The evaluation outcomes of all DL models using the train-test split method. . . . 51
9.3 The proposed model is compared to some of the most recent related research works. . . . 55

Chapter 1

Introduction and Literature Survey

1.1 Introduction

1.1.1 Overview of HIV/AIDS

HIV is a virus that affects cells that help the body resist infection, causing it more resistant to other pathogens and diseases. HIV spreads by contact with the body fluids of a person with HIV, most often through casual intercourse (sex without a condom or HIV medication to avoid or cure HIV), or by the sharing of injection drug supplies [1]. AIDS is the late stage of HIV infection that happens after the body's immune system is severely compromised by the virus. Figure 1.1 shows the structure of HIV virus.

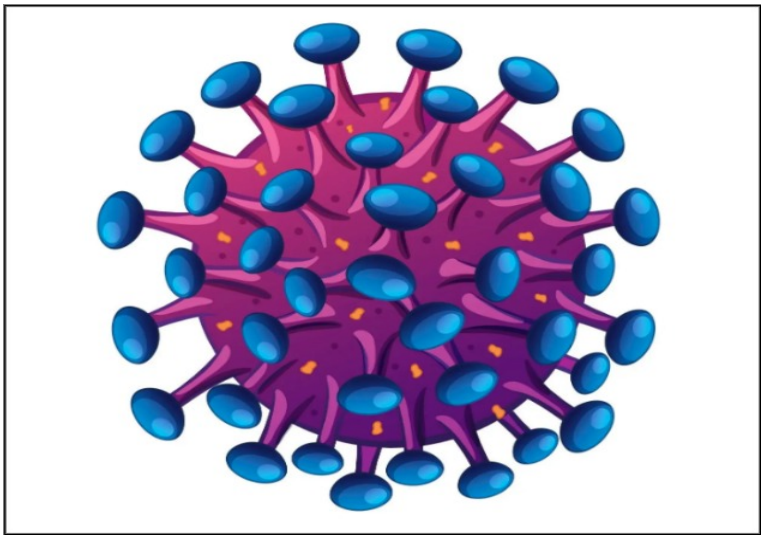

FIGURE 1.1: The HIV virus structure [1]

1.1.2 Status of HIV Epidemic in Ethiopia

The HIV outbreak in Ethiopia is heterogeneous by genders, geographical areas and population groups. HIV incidence in women and men combined is seven times higher in urban areas than in rural areas (2.9 percent versus 0.4 percent). HIV prevalence among urban women is 3.6%, compared with 0.6% among rural women. Seven of the nine provincial states and two city administration have HIV prevalence in excess of 1%. Looking at

HIV prevalence by area, it is highest in Gambella (4.8%), followed by Addis Ababa (3.4%), Dire Dawa (2.5%) and Harari (3.4%). (2.4%) [2].

In Ethiopia, HIV is found in 0.9% of women and men aged 15 to 49. Women are more likely than men to be infected with HIV [2].

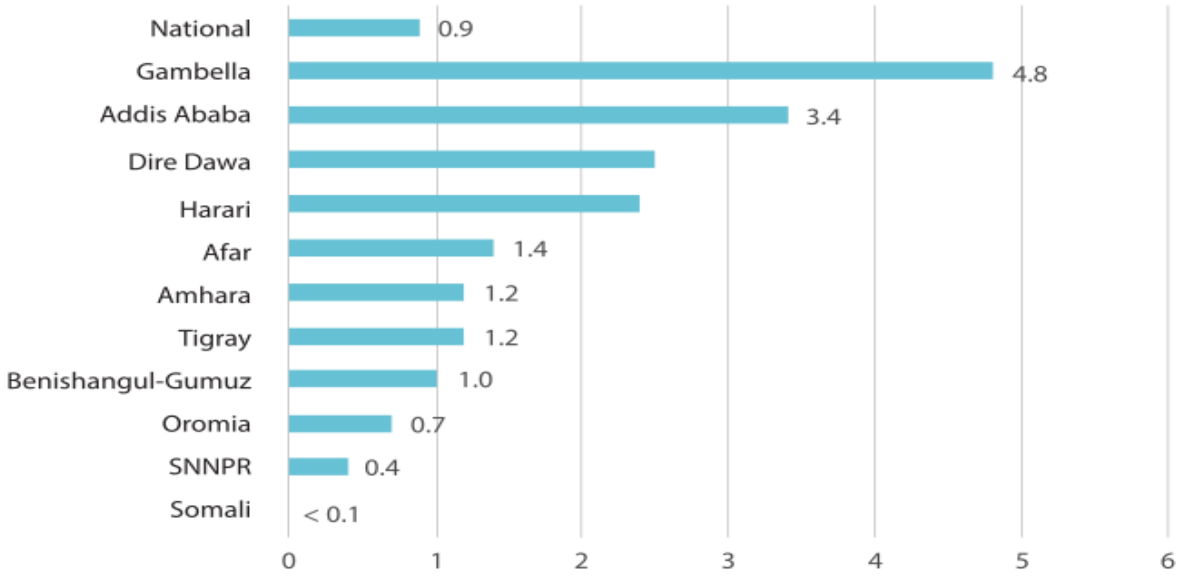

FIGURE 1.2: HIV Prevalence among adult men and women by Region in Ethiopia EDHS 2016 [2]

An estimated 613,000 HIV-positive people, of whom 62% were female, were living in Ethiopia in 2017. Despite the varying prevalence rates in the countries, it is important to consider the absolute number (People living with HIV) of PLHIV per area as the population size varies from one region to another. Amhara, Oromia, and Addis Ababa account for three-quarters (74%) of PLHIV infections [3]. The distribution of PLHIV by area is depicted in the Figure 1.3.

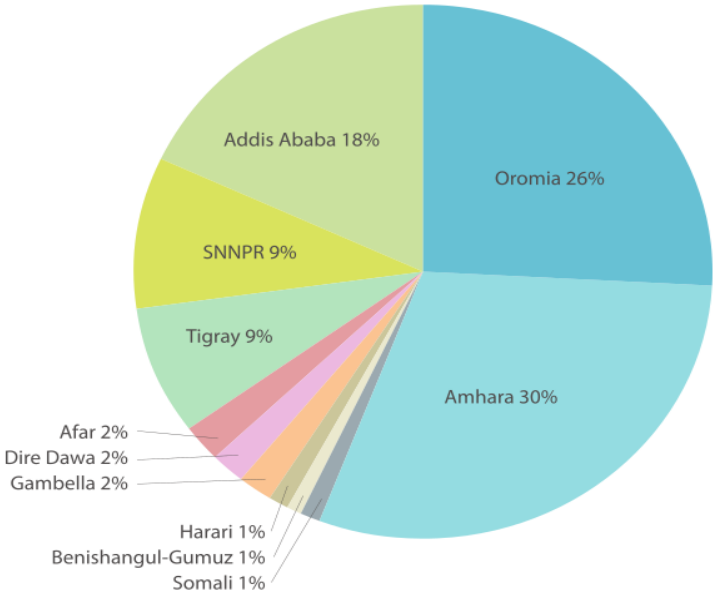

FIGURE 1.3: Proportion of People Living with HIV (PLHIV) by Region in Ethiopia [2]

1.1.3 Ethiopia Demographic and Health Survey (EDHS)

The Ethiopian Demographic and Health Survey (EDHS) was administered by the Central Statistical Agency (CSA) under the auspices of the Ministry of Health. The Ethiopian Health and Nutrition Research Institute (EHNRI) was responsible for HIV tests from dried blood samples (DBS). Demographic and Health Survey (DHS) undertaken in Ethiopia under the global MEASURE DHS program, a USAID-funded project offering funding and technical assistance in the implementation of population and health surveys in countries around the world [4].

The EDHS' main goals are to provide up-to-date data for population and health service preparation, policy development, reporting, and assessment.

For the nine regional states and two city governments, the EDHS study includes core indicators on family planning, fertility levels and determinants, fertility preferences, infant, child, adult, and maternal mortality, maternal and child welfare, diet, women's empowerment, and HIV/AIDS awareness. In addition, at the country level, this study includes details by urban and rural residence [4].

This study uses EDHS as a source of data, more specifically the HIV/AIDS dataset. The detail explanation of the EDHS is presented in Chapter 2.

1.1.4 Literature Review

There are large number of attempts have been done by different researchers for feature selection, classification and prediction using various techniques on health data. For the purpose of this work we tried to put some of the research works which is related to our methodology as shown in Table 1.1.

TABLE 1.1: Recent papers from standard publications are surveyed, reviewed and summarized in the table below.

Begin of Table						
No.	**Year**	**Authors**	**Title**	**Applications/Focus of work**	**Methods**	**Dataset**
1.	2015	Pinar Yildirim [5]	Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease	Prediction of Risks in Hepatitis Disease	Information Gain, Relief, One-R, PCA, CFS, and Consistency Based Subset Evaluation	Hepatitis, UCI repository
2.	2011	Karegowda, A. G., et al.[6]	Feature Subset Selection using Cascaded GA and CFS: A Filter Approach in Supervised Learning	To evaluate the effect GA and CFS	The Fitness Function and Genetic Algorithms	PIDD from UCI
3.	2013	Sarkar, S. and Goswami, S. [7]	Empirical Study on Filter based Feature Selection Methods for Text Classification	Comparing performance of feature selection techniques employed with different classifiers	Chi-squared, Information gain, Mutual Information, and Symmetrical Uncertainty	CNAE-9, SMS spam collection, and Preview of hotel data
4.	2019	Rajit Nair and Amit Bhagat [8]	Feature Selection Method To Improve The Accuracy of Classification Algorithm	to shows how feature selection works and how it helps for classification process based on selectkbest and selectpercentile.	Univariate method, Model based feature selection, and Iterative selection	Breast Cancer Dataset
5.	2014	Harb, H. and Desuky, A. [9]	Feature Selection on Classification of Medical Datasets based on Particle Swarm Optimization	Perform the comparison of filter and wrapper Methods with PSO to improve the classifiers performance	PSO+NBay, PSO+CFS and GA+CFS	Breast Cancer Dataset
6.	2013	Khan, M. and Quadri, S.[10]	Effects of Using Filter Based Feature Selection on the Performance of Machine Learners Using Different Data sets	Reviews different feature selection methods: filter, wrapper and embedded, that help in selecting the optimal feature subsets	Univariate Filter, Multivariate Filter, Deterministic Wrapper, Randomized Wrapper, and Embedded	Australian Credit Approval dataset from UCI
7.	2015	Chitra, D and Nasira, G[11]	Wrapper Based Feature Selection For CT-Image	Investigate various feature selection algorithms for classification of CT-Image	Information gain, Bacterial Foraging Optimization, and Correlation based Feature Selection	Medical image data

Continuation of Table 1.1

No.	Year	Authors	Title	Applications/Focus of work	Methods	Dataset
8.	2015	Panthong, R. and Srivihok, A.[12]	Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm	To evaluate a subset of features	Sequential forward selection (SFS), sequential backward selection (SBS) and optimize selection (evolutionary)	13 datasets from the UCI
9.	2010	Karegowda, A. G, et al.[13]	Feature Subset Selection Problem using Wrapper Approach in Supervised Learning	Evaluate features to enhance the classification accuracy	Genetic Algorithm	PIMA
10.	2012	Ozcift, A. and Gulten, A.[14]	A Robust Multi-Class Feature Selection Strategy Based on Rotation Forest Ensemble Algorithm for Diagnosis of Erythemato-Squamous Diseases	Evaluate the performance of Rotation Forest Ensemble Feature Selection approach	Rotation Forest Ensemble	Erythemato-Squamous diseases dataset
11.	2018	Abinash, M. and Vasudevan, V.[15]	A Study on Wrapper-Based Feature Selection Algorithm for Leukemia Dataset	Discusses the SVM-based wrapper feature selection for cancer classification	Correlation and SVM	Leukemia
12.	2005	Talavera, L.[16]	An Evaluation of Filter and Wrapper Methods for Feature Selection in Categorical Clustering	Implementation of a wrapper and adapt an existing filter method	Wrapper and Filter	10 dataset from UCI
13.	2014	Nnamoko, N. et al. [17]	Evaluation of Filter and Wrapper Methods for Feature Selection in Supervised Machine Learning	Examines adapting an existing wrapper and filter method to perform experiments with a diabetes diagnosis	Information Gain For filter and Classifier SubsetEval For wrapper	Diabetes From UCI
14.	2013	Cadenas, J. et al [18]	Feature subset selection Filter-Wrapper based on low quality data	To improve the classification performance of features based on the Fuzzy Random Forest	Fuzzy Random Forest For both Filter and Wrapper	10 microarray datasets

Continuation of Table 1.1

No.	Year	Authors	Title	Applications/Focus of work	Methods	Dataset
15.	2010	Min, H., and Fangfang, W.[19]	Filter-Wrapper Hybrid Method on Feature Selection	To optimize the efficiency of features	For filter: Correlative Family Selection, Relief, Class Separability, Mahalanobis Distance, Multivariate Correlation Coefficient and Mutual Information. For wrapper: Adaptive Genetic Algorithm, Chaotic Binary Particle Swarm Optimization and Clonal Selection Algorithm	Yangtze River Tunnel engineering
16.	2013	Yang, P. et al.[20]	Ensemble-Based Wrapper Methods for Feature Selection and Class Imbalance Learning	To create multiple balanced datasets from the original imbalanced dataset via sampling, and subsequently evaluate feature subsets	Ensemble-based wrapper	5 leukemia dataset
17.	2017	Fatima, M. and Pasha, M. [21]	Survey of machine learning algorithms for disease diagnostic	To provide the comparative analysis of different machine learning algorithms for diagnosis of different diseases	Machine Learning	Cleveland heart dataset from UCI
18.	2018	Idowu, P. et al [22]	Prediction of pediatric HIV/AIDS survival in Nigeria using Naive Bayes' Approach	To predict the survival of patients from HIV/AIDS	10 fold cross validation with Statistical methods	216 pediatric HIV/AIDS Patients
19.	2016	Koçak, Y. et al.[23]	Classification of HIV data by constructing a social network with frequent itemsets	Proposed a novel classification algorithm to classify HIV	Frequent Item set Network Based Classifier	HIV-1 protease cleavage data set from UCI
20.	2016	Olayemi Olufunke et al.[24]	Predictive Model of Pediatric HIV/AIDS Survival in Nigeria using Support Vector Machine	To develop a predictive model for the classification of HIV/AIDS survival	Sequential minimal optimization	HIV/AIDS pediatric patients in Nigeria

Continuation of Table 1.1

No.	Year	Authors	Title	Applications/Focus of work	Methods	Dataset
21.	2017	Oliveira, A. et al [25]	Data Mining in HIV-AIDS Surveillance System	To identify the main factors influencing reporting delays of HIV-AIDS cases	Machine Learning	Portuguese HIV-AIDS data
22.	2010	WU, H. et al.[26]	A HIV Carrier Forecasting Model for Quarantine Based on Support Vector Machines	To construct a HIV carrier forecasting model based on support vector machines	SVM	Port of a province in China during the period of 2004-2009
23.	2005	Leke-Betechuoh, B. et al.[27]	Prediction of HIV status from demographic data using neural networks	To classify HIV status of individuals based on demographic and socioeconomic characteristics.	RBF	South African antenatal seroprevalence survey of 2001
24.	2018	Igulot, P. and Magadi, M.[28]	Socioeconomic status and vulnerability to HIV infection in Uganda: Evidence from multilevel modelling of AIDS indicator survey data	To examine the association between socioeconomic status (SES) and HIV infection in Uganda and to examine whether the SES-HIV relationship	Multilevel logistic regression	2011 Uganda AIDS
25.	2020	Anisha, C. and Arulanand, N[29]	Early Prediction of Parkinson's Disease (PD) Using Ensemble Classifiers	Prediction of PD based on speech symptoms	Ensemble Classifier PCA, LDA	PD dataset from UCI
26.	2019	Shekar, B. H. and Dagnew, G.[30]	Grid search-based hyperparameter tuning and classification of microarray cancer data	To classify microarray cancer data	Gini index	5 standard microarray Cancer Dataset
27.	2016	Mantovani, R. G. et al.[31]	Hyper-parameter tuning of a decision tree induction algorithm	Investigate DT induction algorithms to HP tuning process	Random Search strategy (RS), meta-heuristics (MTHs) - Genetic Algorithms (GAs),Particle Swarm Optimization (PSO), and Estimation of Distribution Algorithm (EDA)	UCI repository
28.	2019	Probst, P. et al.[32]	Tunability: Importance of hyperparameters of machine learning algorithms	To assess the tunability of the parameters.	Machine Learning	38 datasets from the OpenML

Continuation of Table 1.1

No.	Year	Authors	Title	Applications/Focus of work	Methods	Dataset
29.	2019	Ahlström, M. et al. [33]	Algorithmic prediction of HIV status using nation-wide electronic registry data	To predict HIV status using machine learning algorithms	Machine Learning	Danish National Hospital Registry
30.	2018	Annang, D. A.[34]	Performance Comparison of Data Mining Techniques for Predicting Hiv Status Among Female Sex Workers in Ghana	To report summaries of HIV status of FSW with their respective socio-demographic and behavioral variables using traditional statistical methods.	CRISP-DM	2015 IBBSS for FSW
31.	2015	Hailu, T. G.[35]	Comparing data mining techniques in HIV testing prediction	Build the model that predicts whether an individual was being tested for HIV among adults in Ethiopia using EDHS 2011.	Cross Industry Standard Process Data Mining (CRISP-DM)	2011 EDHS dataset
32.	2012	Girma Awoke [36]	Predicting HIV Infection risk factor using voluntary counseling and and testing data	Investigate the underlying determinant risk factors of HIV infection from the available VCT data using data mining techniques	CRISP-DM	African AIDS Initiative International, Voluntary HIV/AIDS Counseling and Testing (AAII VCT)
33.	2010	Teklu Urgesa [37]	Application of data mining techniques on antiretroviral therapy (ART) data	Identifying the determinant factors affecting the termination/continuation of the service	CRISP-DM	Adama and Asella Hospitals ART databases
34.	2016	Birhanu Fetene [38]	Application of data mining for effective prognosis of HIV/AIDS testing	Prognosis of HIV/AIDS testing	Hybrid methodology CRISP-DM and KDD	EDHS 2011
35.	2011	Elias Lemuye [39]	HIV status Predictive Modeling using data mining technology	Investigate the underluing determinant factor of being HIV positive or HIV negative from the Available HCT data using datamining technology	CRISP-DM	HCT data from
36.	2020	Yin, B, M. et al. [40]	Predicting the diagnosis of HIV and sexually transmitted infections among men who have sex with men using machine learning pproaches	To develop machine learning models and evaluate their performance in predicting HIV STI	Machine Learning	MSM and STI from 2011-2017

Continuation of Table 1.1

No.	Year	Authors	Title	Applications/Focus of work	Methods	Dataset
37.	2020	McSharry, P. E., et al. [41]	Use of machine learning techniques to identify hiv predictors for screening	Analysing the HIV disease trial at various levels of society, detect HIV predictors, and forecast the risk of the disease	Machine Learning	Population-based HIV Impact Assessment
38.	2020	Orel, E., et al. [42]	Machine learning to identify socio-behavioural predictors of hiv positivity in east and southern africa	Examined morethan 3,200 parameters of the current Demographic HealthSurveys from ten African nations	Machine Learning	DHS data
39.	2019	Ahlström, M. G., et al. [33]	Algorithmic prediction of hiv status using nation-wide electronic registry data	Investigated whether data from a national electronic registry might be utilised to predict HIV status	Machine learning	Danish National Hospital Registry UCI ML repository
40.	2018	Lu, X., et al. [43]	The application of deep learning in the prediction of hiv-1 protease cleavage site	Employed one-hot coding to translate protease cleavage site data for prediction	Deep Learning	UCI ML repository
41.	2020	Steiner, M. C., et al. [44]	Drug resistance prediction using deep learning techniques on hiv-1 sequence data	Evaluate the DL models for drug resistance prediction using HIV-1	Deep Learning	HIV Drug Resistance database
42.	2019	Wang, G., et al. [45]	Application of a long short term memory neural network: a burgeoning method of deep learning in forecasting hiv incidence in guangxi, china	created convenient model to explain the prevalence of HIVand forecast its occurrence	Deep Learning	HIV incidence statistics data

End of Table

1.2 Motivation

The Ethiopian Demographic and Health Survey (EDHS) dataset provides massive amounts of information about the characteristics of HIV/AIDS and HIV monitoring. Predictive patterns in this data set cannot be discovered using standard and well-known statistical methods or applications. It is therefore of concern to academics, policymakers, and HIV program administrators to be able to turn data from surveys such as the EDHS into a models for analyzing an individual's HIV status. This would assist in the transformation of academic discoveries into realistic results that can be incorporated in the real world. As Ethiopia, the Ministry of Health (MOE), and other authorised organisations continue to develop interventions to improve people's use of HIV testing services, predictive methods like this one will supplement efforts on successful HIV programming and planning. This is due to the fact that if the HIV status of individuals (people) can be correctly estimated by a certain proportion, it would also help address concerns about HIV testing among people.

The underlying issue that prompted this study is the ever-expanding pandemic of HIV/AIDS in the world in general, and in Ethiopia in particular, as well as the complexities of preventing and monitoring it. The problem deserves careful thought and has become a topic of study for many academics. Different researches have been done on HIV/AIDS prediction in Ethiopia and abroad with small proportion of the dataset and measure the performance of the model only based on accuracy and also they used only percentage split method to classify the dataset into training and testing dataset. However, since all previous studies are performed using a CRISP-DM Methodology, this does not fully address the issue that was encountered during HIV/AIDS prognosis.

Few number of study has been conducted in Ethiopia to map the magnitude of the HIV/AIDS problem using data mining technologies. All of the previous studies have followed the same methodologies and they have limited in their study area/scope. Biru [46] proposes the applicability of data mining techniques to support voluntary counseling and testing (VCT) for HIV to investigate the potential applicability of data mining techniques on VCT client dataset in developing a model that could support VCT in identifying patterns. Teklu [37] proposes the application of data mining techniques on antiretroviral Therapy (art) data: The case of adama and asella hospitals to identifying the determinant factors affecting the termination /continuation of the service. Girma [36] Predicting HIV infection risk factor Using voluntary counseling and testing data to investigate the underlying determinant risk factors of HIV infection from the available VCT data using data mining techniques. Another specific studies has also conducted in the prediction of HIV status, Elias [39] studies the HIV status predictive modeling using Data mining technology to investigate the underlying determinant factor of being HIV positive or HIV negative. Tesfaye [35] comparing data mining techniques in HIV testing prediction to predict the test results among adults. Birhanu [38] apply data mining for effective prognosis of HIV/AIDS testing to predict HIV.

To the knowledge of the researcher, no attempt has been made using machine learning methodologies with EDHS-HIV data set in Ethiopia. Moreover, this work has been the first attempt to explore and evaluate the feature selection techniques and proposes a machine learning models for classification and prediction of the HIV/AIDS status using EDHS dataset. Therefore, the purpose of this study is to explore and analyses the predictive models of machine learning to predict the status of HIV epidemics.

1.3 Objectives of the Research

The following is the list of objectives for exploring predictive models for analysis of the epidemic status of HIV/AIDS:

1. To apply preprocessing tasks to make the dataset suitable for the experiments.
2. To explore feature selection methods and apply their approach.
3. To examine a classification algorithm for analysis of the epidemic status of HIV/AIDS.
4. To develop a grid search method for HIV/AIDS data analysis.
5. To deploy multiple deep learning models for HIV status prediction.

1.4 Justification of the Research

The knowledge of the prevalence of HIV/AIDS disease helps in providing information on designing appropriate control and preventive integrative measures in order to achieve a long-term solution. The research will be helpful in order to recognize the status and predict the prevalence of HIV/AIDS so that it gives insight to take precautionary measures. It also guides policymakers in making appropriate interventions to control and minimise the consequences of HIV/AIDS.

Chapter 2

Research Methodology

2.1 Dataset Collection and Pre-processing

2.1.1 Data Source

This study was based on secondary data from the DHS program https:www.dhsprogram.com. DHS facilitates analysis of Demographic and Health Surveys, administered in low and middle income countries since the 1980s. DHS contains thousands of consistently coded variables on the health and well-being of women, children, births, men, and on all members of randomly selected households, for 32 African countries and 9 Asian countries. We extract Ethiopia's DHS dataset from the 2000, 2005, 2011, and 2016 EDHS database. The EDHS included a nationally representative group of women (aged 15 to 49) and men (aged 15 to 59) from each of the country's eleven administrative regions [4].

The following Table 2.1 shows the EDHS dataset which is collected from the Four separate survey. Each dataset has consists of two separate datasets which are the Male records and Women records. The data base was manipulated by using SPSS software system.

TABLE 2.1: The Total records of the EDHS dataset of four separate year of Survey

No.	Year of Survey	Variable		Instances		Total records	
		Men	Women	Men	Women	Variables	Instances
1.	2000	491	3,664	4,772	5,014	4,155	9,786
2.	2005	960	4.699	5,248	9,070	5,659	14,318
3.	2011	547	3,807	14,110	16,515	4,354	30,625
4.	2016	590	5,704	12,688	15,683	6,294	28,371
					Total	**20,462**	**83,100**

For the purpose of this work, we select only the HIV/AIDS related records or information from the initial EDHS dataset manually and ready for the feature selection methods. Table 2.2 Presented the initial HIV/AIDS records which is manually extracted from the EDHS dataset from each years of survey.

TABLE 2.2: The Initial records of the HIV/AIDS dataset of Four Separate year from EDHS Survey

No.	Year of Survey	Variables for Men and Women	Instances		Total Records	
			Men	Women	Variables	Instances
1.	2000	33	4,772	5,014	33	9,786
2.	2005	33	5,248	9,070	33	14,318
3.	2011	33	14,110	16,515	33	30,625
4.	2016	33	12,688	15,683	33	28,371
		Total	**36,818**	**46,282**	-	**83,100**

2.1.2 Data Preprocessing

2.1.2.1 Data Integration

Data integration means integrating data from a number of different sources, which are processed using different methods to have a single view of the data. There are primarily two major approaches to data integration – one is a "tight coupling approach" and the other is a "loose coupling approach" [47]. We apply the "tight coupling technique" In this step, HIV/AIDS data is combined from various sources, from four independent EDHS databases to a single physical location via the (ETL-Extraction, Transformation and Loading) process. Based on these approach, we combined all separate dataset from EDHS as one dataset. The total dat set we conduct after integration is 83,100 instances and 33 variables. The data integration is done by GNU pspp version 1.0.1 tool, it is a program for the analysis of sampled data.

2.1.2.2 Data Cleaning

The problem like user entry error, storage or transmission error, or by various data dictionary definition of similar data in different stores. These data are unreliable, inaccurate or irrelevant for HIV/AIDS test result prediction using machine learning models. To treat these problem, we apply data cleaning techniques to identify and removing the crude and incorrect instances. There are different types of techniques to clean those data with problems. The data cleaning technique applied in this process is remove duplicate technique and delete all formatting technique. From the total 33 variables of the original features of the HIV/AIDS dataset, we remove “Age_of_House_Hold”, “Sex_of_House_Hold”, “Occupation”, “Relationship_to_House_Hold”, “Literacy”, and “Ethnicity” based on the concepts of the data cleaning techniques used in this process. After removing and deleting the unreliable, and irrelevant features with their values, we fond 26 features with their values for future processing.

2.1.2.3 Handling Missing Value

We reviewed the dataset for HIV testing and found missing values and took action to address the missing problem. The missing values for the numeric data type and the usually distributed data were replaced by the variable mean. In the case of a categorical variable, the missing values is replaced by the modal value of the variable. The missing values were explicitly defined and the proportion of the amount was determined against the overall records. As a result, all the selected attributes have less than five percent missing values 5%. The selected variables “Recent_sexual_Activity”, “Ever_Had_Sex”, “Condom_Used” are counted for 4.6%, 3.9%, and 3.36% respectively.

2.1.3 Data Transformation

2.1.3.1 Data Discretization

In HIV/AIDS dataset, the multiple value features or variables such as “AGE”, “Region”, “ Educational Level”, “Educational Attainment”, are discretized with explicit data grouping technique. This indicates that the detailed definitions have been grouped into the appropriate level of a more general definition. Almost all of the chosen attributes have been transformed from their original state in such a manner that they can be readily understood and interpreted. The attribute "Religion" used to have six different values (Orthodox, Traditional, Protestant, Catholic, Muslim, and others), but it was later split into three categories: Christian (Orthodox, Protestant, and Catholic), Muslim, and others.

2.1.3.2 Data Reduction

In this research we applied the lossless which does not lose any data in the compression process.

Table 2.3 portrayed the detail of re-coding with regard to feature/variables name, their existing value and their new numeric value as well.

The conversation is performed manually, we simply re-code the nominal values to numeric values. Those feature/variable values are changed to numeric format and followed with data size reduction. The encoding techniques handled with SPSS 20, Transform menu of re-code into same variable futures; this enables to change the existing feature/variable values with the given numeric value without creating additional variables.

Out of 26 features from HIV/AIDS data set the values of the 25 features are encoding or changed from nominal values to numerical manually using SPSS 20 tool. Only the value of “Age” variables are numeric from the entire features.

2.1.3.3 Data Normalization

Normalization is a technique for reducing the variance of an attribute’s data to a narrower range, such as -1.0 to 1.0 or 0.0 to 1.0. When working with attributes on various scales, normalisation is usually required; otherwise, the utility of an important, similarly important attribute (on a lower scale) can be diluted due to the presence of other attributes with values on a larger scale. Simply put, where multiple attributes exist but their values are on different scales, weak data models can result when conducting data mining operations [48].

In this research work, we used “Min-Max Normalization” techniques. The initial data is transformed linearly in the Min-Max data normalisation technique. The minimum and maximum values from the data are retrieved, and each value is replaced using the formula below.

$$V = \frac{v - min(A)}{max(A) - min(A)} \tag{2.1}$$

where A denotes the attribute data, and Min(A) and Max(A) denote the attribute data’s minimum and maximum absolute values, respectively. Each data entry’s new value is denoted by the letter V. v is the previous value of each data entry [49].

TABLE 2.3: Encoding new value for replacement of existing value

S.No.	Feature Name	Existing Values	New Values
1.	Sex	Male, Female	1,2
2.	Region	Tigray, Afar, Amhara, Oromia, Somali, Benshangul, SNNP, Gambela, Harari, Addis Ababa, Diredawa	1,2,3,4,5,6,7, 8,9,10,11
3.	Resident_Place	Urban, Rural	1,2
4.	Relegion	Chrstian, Muslim, Others	1,2,3
5.	Educational_Level	No education, Primary, Secondary, Higher	0,1,2,3
6.	Educational_Attainment	No education, Incomplate Primary, complate Primary, Incomplate Secondary, Complate Secondary, Higher	0,1,2,3,4
7.	Martial_Status	Never married, Married, Living Togather, Widowed, Divorced, Not living togather	0,1,2,3,4,5
8.	Currently_working	No, Yes	0,1
9.	Welth_Index	Poorset, poorer, Middle, Richer, Richest	1,2,3,4,5
10.	Recent_sexual_activity	Active, Not active	1,0
11.	Number_of_sex_partner	One, Morethan One	0,1
12.	Ever_had_sex	Yes, No	1,0
13.	Condom_used	Yes, No	1,0
14.	Reduced_HIV_use_condom	Yes, No	1,0
15.	Reduced_HIV_1Sex_partner	Yes, No	1,0
16.	Reduce_HIV_DoNot_Use_ Condom_At_ALL	Yes, No	1,0
17.	Can_get_HIV_ by_Mosquito	Yes, No	1,0
18.	Heard_STI	Yes, No	1,0
19.	Heard_Other_STI	Yes, No	1,0
20.	Ever_Heared_HIV	Yes, No	1,0
21.	Ever_tested_HIV	Yes, No	1,0
22.	Know_Place_of_HIV_test	Yes, No	1,0
23.	Sample_Test	Yes, No	1,0
24.	Test_in_LAB	Yes, No	1,0
25.	Final_Test_result	Positive, Negative	1,0

Chapter 3

Feature Selection Methods for Prediction of the Individual's Status of HIV/AIDS from EDHS Dataset - A Filter Approach

3.1 Introduction

In this study, the high interest of HIV/AIDS disease, which is a major health issue around the world was considered and we explore and examine the existing filter based feature selection methods namely (I) Univariate Chi-squared (II) Feature Importance and (III) Correlation coefficient.

The primary aim of this study is to make a contribution to HIV/AIDS status prediction for health research by presenting a systematic and quantitative comparison of existing filter-based feature selection methods.

3.2 Filter Based Feature Selection Methods

3.2.1 Univariate Feature Selection

In this paper we select best features from the total original feature using chi-square methods. Because chi-square metrics is a model agnostics, rely entirely on features, and computationally very fast. The main justification for selecting this metrics is: the dataset used in the study is discrete and categorical.

Chi is a statistical test which measures a feature's independence from the class labels. It is a bidirectional metric. As Forman [50] noted that this test can act erratically when feature counts are low to be predicted. With class imbalanced data sets this approach is relatively popular. In this technique we divide the features with **10**, **15**, and **20** randomly, and we select best **20** features among the total of 26 features from EDHS-HIV/AIDS data set. Table 3.1, presents the selected features using chi square technique of univariate methods.

3.2.2 Feature Importance

Figure 3.1 shows how feature choices are made using feature importance selection. We use Extra Tree classifier model to extract the top **20**, top **15**, and top **10** features from the EDHS-HIV/AIDS dataset. The selected features are selected based on the scores how they are more useful to predicting the target class F_T_Resu.

TABLE 3.1: Selected features using univariate feature selection methods

Rank	Specs	Score
18	H_STI	4826.220044
23	S_Test	4653.720019
24	T_in_LAB	1087.105393
2	Reg	933.170027
22	P_T_HIV	879.389418
17	HIV_Mosq	465.323441
0	Sex	447.536106
11	N_S_Part	345.145407
6	Edu_At	258.215903
5	Edu_lvl	147.122366
7	M_Sta	121.090588
19	H_O_STI	119.945452
21	E_T_HIV	80.808937
10	R_SeA	57.726482
1	Age	32.821387
14	R_Use_Con	31.845463
12	Had_Sex	22.130725
20	H_AIDS	11.966975
4	Rel	11.696209
8	C_Wor	9.172250

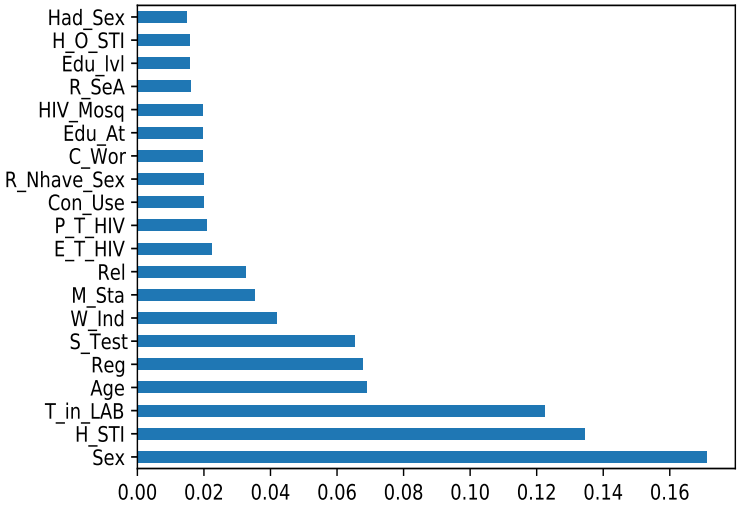

FIGURE 3.1: Selected features using feature importance criteria

3.2.3 Correlation Coefficient

In this paper, we used correlation coefficient with Heat-map which makes it easy to identify which features are most closely related to the target variable. Figure 3.2 shows how the target variables correlate the features. The results of CC's metrics as depicted from Figure 3.2, the feature H_STI is highly correlated with F_T_Resu followed by S_Test and HIV_Mosq. Con_Use and Res_P are the least correlated features with the target features.

3.3 Experimental Results

Table 3.2 shows the experimental results of the proposed method. The highest accuracy values are obtained for the HIV/AIDS data set using Univariate Feature selection with Random forest, Gradient boosting, K-nearest

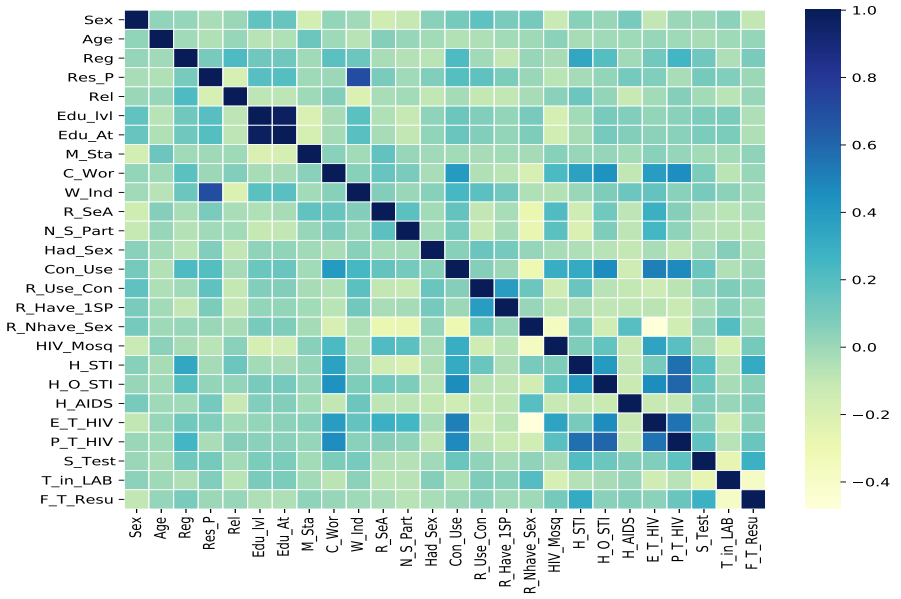

FIGURE 3.2: Heatmap of Pearson correlation coefficient matrix for the considered EDHS dataset

neighbors, Support vector machine, Ada-boost, Naive bayes and Logistic regression classifiers are **0.95**, **0.91**, **0.90**, **0.83**, **0.81**, **0.76** and **0.76** respectively. The above scores are obtained from the selected feature with **20** features from the total features compared with selected features of univariate such as feature **15** and **10**.

Similarly, in the Feature Importance Feature selection methods, the experimental result is obtained with same classifiers which are used in univariate methods and the result with Random forest, Gradient boosting, K-nearest neighbors, Ada-boost, Support vector machine, Logistic regression and Naive bayes are **0.92**, **0.91**, **0.90**, **0.82**, **0.79**, **0.78** and **0.72** respectively.

In this Experiments, we also test Correlation coefficient based feature selection with **9** best features which are highly correlated with the target value. The experimental result are obtained with the same classifier Gradient boosting, Random forest, K-nearest neighbors, Support vector machine, Ada-boost, Naive bayes and Logistic regression is **0.88**, **0.86**, **0.84**, **0.83**, **0.79**, **0.74** and **0.74** respectively. In addition to the above evaluation metrics, we used ROC as a parameter to evaluate the selected features. The ROC result of each selected features with the same classifiers obtained to compare one to the others. As we observe from the experiment, the filter based feature selection with univariate with the chi-square criteria is best for selecting or choosing the best features with compared to the Feature importance and correlation coefficient.

TABLE 3.2: Result of filter based feature selection methods on EDHS-HIV/AIDS dataset

Classification Algorithms	**Feature Selection Methods**	**No. of Selected Features**	**Evaluation Metrics**				
			Acc	**Pre**	**Rec**	**F1-score**	**ROC**
Random Forest (RF)	**Univariate**	**20**	**0.952**	**0.925**	**0.984**	**0.954**	**0.95**
		15	0.890	0.842	0.962	0.898	0.89
		10	0.890	0.842	0.962	0.898	0.89
	Feature Importance	**20**	**0.952**	**0.885**	**0.970**	**0.925**	**0.92**
		15	0.922	0.885	0.970	0.925	0.92
		10	0.869	0.817	0.950	0.879	0.87

Continued on next page

Table 3.2 – Continued from previous page

Classification Algorithms	Feature Selection Methods	No. of Selected Features	Evaluation Metrics				
			Acc	Pre	Rec	F1-score	ROC
	Correlation Coefficient	**9**	0.878	0.889	0.942	0.915	0.88
K-Nearest Neighbour (KNN)	**Univariate**	**20**	**0.901**	**0.871**	**0.942**	**0.905**	**0.90**
		15	0.866	0.827	0.927	0.874	0.87
		10	0.828	0.827	0.833	0.830	0.83
	Feature Importance	**20**	**0.900**	**0.868**	**0.943**	**0.904**	**0.90**
		15	0.883	0.848	0.932	0.888	0.88
		10	0.848	0.806	0.917	0.858	0.85
	Correlation Coefficient	**9**	0.844	0.848	0.844	0.846	0.84
Support Vector Machine (SVM)	**Univariate**	**20**	**0.833**	**0.776**	**0.937**	**0.849**	**0.83**
		15	0.823	0.775	0.911	0.838	0.82
		10	0.821	0.773	0.908	0.835	0.82
	Feature Importance	**20**	**0.791**	**0.733**	**0.917**	**0.814**	**0.79**
		15	0.783	0.716	0.940	0.812	0.78
		10	0.821	0.773	0.909	0.836	0.82
	Correlation Coefficient	**9**	0.830	0.774	0.931	0.845	0.83
Naive Bayes (NB)	**Univariate**	**20**	**0.760**	**0.741**	**0.804**	**0.771**	**0.76**
		15	0.732	0.692	0.845	0.761	0.73
		10	0.730	0.690	0.845	0.759	0.73
	Feature Importance	**20**	**0.720**	**0.676**	**0.856**	**0.755**	**0.72**
		15	0.709	0.660	0.872	0.751	0.71
		10	0.709	0.660	0.872	0.751	0.71
	Correlation Coefficient	**9**	0.738	0.707	0.820	0.760	0.74
Logistic Regression (LR)	**Univariate**	**20**	**0.758**	**0.743**	**0.801**	**0.771**	**0.76**
		15	0.732	0.696	0.842	0.762	0.73
		10	0.731	0.690	0.844	0.759	0.73
	Feature Importance	**20**	**0.779**	**0.805**	**0.903**	**0.851**	**0.78**
		15	0.700	0.940	0.913	0.753	0.70
		10	0.700	0.636	0.933	0.757	0.70
	Correlation Coefficient	**9**	0.738	0.711	0.819	0.761	0.74
AdaBoost (AB)	**Univariate**	**20**	**0.811**	**0.790**	**0.854**	**0.821**	**0.81**
		15	0.782	0.773	0.806	0.789	0.78
		10	0.770	0.770	0.774	0.772	0.77
	Feature Importance	**20**	**0.817**	**0.793**	**0.961**	**0.826**	0.82
		15	0.798	0.774	0.841	0.806	0.80
		10	0.776	0.774	0.787	0.780	0.78

Continued on next page

Table 3.2 – Continued from previous page

Classification Algorithms	Feature Selection Methods	No. of Selected Features	Evaluation Metrics				
			Acc	Pre	Rec	F1-score	ROC
	Correlation Coefficient	**9**	0.786	0.783	0.803	0.793	0.79
Gradient Boosting (GB)	**Univariate**	**20**	**0.905**	**0.906**	**0.907**	**0.907**	**0.91**
		15	0.880	0.887	0.887	0.882	0.88
		10	0.851	0.876	0.819	0.847	0.85
	Feature Importance	**20**	**0.905**	**0.904**	**0.908**	**0.906**	**0.91**
		15	0.897	0.905	0.890	0.897	0.90
		10	0.876	0.894	0.855	0.874	0.88
	Correlation Coefficient	**9**	0.856	0.879	0.825	0.852	0.86

3.4 Conclusion

In this research, a proposed feature selection approach was conducted using filter-based feature selection methods to predict the individual status or test outcome of HIV / AIDS from the EDHS data set. The research uses three methods of selecting features under the filter base to classify the data set and assess their output using Random forest, K-nearest neighbors, Support vector machine, Gradient boosting, Adaboost, Naive bayes and Logistic regression. The performance was measured by five evaluation metrics namely: Accuracy, Precision, Recall, F1-measure and ROC. From the experiment, Random forest, K-nearest neighbors and Gradient boosting classifiers have higher accuracy levels on EDHS-HIV/AIDS data set than the others after the applying the filter based feature selection methods. This study shows that methods of selecting features are capable of improving the learning algorithms efficiency.

Finally, the output of this study can make significant contributions in the prediction of the status of HIV/AIDS result of individuals in health domain research and provide filter based feature selection methods for machine learning studies. As a future work, a research will be designed as a potential job to explore the other methods of selecting features which to compete with filter based on the efficiency of feature selection methods and classification accuracy.

Chapter 4

Wrapper Based Feature Selection Techniques on EDHS-HIV/AIDS Dataset

4.1 Introduction

This section presents wrapper based feature selection methods namely (I) Sequential forward selection (SFS), (II) Sequential backward selection (SBS) and (III) Recursive feature elimination (RFS).

4.1.1 Forward Feature Selection (FFS):

This algorithm begins its search with an empty set, then searches for the feature which helps to achieve the highest classification accuracy, when found, this attribute is simply applied to the empty set which forms the subset of the searched element. This process is replicated as needed until the classification precision can not be further enhanced by including all the remaining elements [51].

For the purpose of this study we are select **20**,**15** and **10** features from EDHS data set using FFS. And here we presents the selected 20 features namely: Sex, Reg, Res_P, Rel, C_Wor, R_SeA, N_S_Part, Had_Sex, Con_Use, R_Use_Con, R_Have_1SP, R_Nhave_Sex, HIV_Mosq, H_STI, H_O_STI, H_AIDS, E_T_HIV, P_T_HIV, S_T est, T_in_LAB.

4.1.2 Backward Feature Selection (BFS):

This approach works similarly to the FFS, but the operation goes in a certain direction. In reality, the algorithm begins with in a group containing all the accessible features, instead it mostly excludes the removal of the feature which improves the accuracy of the classification. This task is replicated again until no classification efficiency can be increased by deleting all of the remaining characteristics [51]. The features are selected after applying this methods from the original data set.

4.1.3 Recursive Feature Selection (RFS):

Recursive elimination of features performs a greedy quest to determine the highest performing subset of features. It iteratively generates models and each iteration decides the best or worst performing feature. This builds the

subsequent models with the left features before it examines all features. This then lists the features according to the order of their elimination. In the worst case, if a set of data includes N number of features RFE can perform a greedy search for 2N feature combinations.

4.2 Experimental Results

Table 4.1 shows experimental results of the proposed method. The highest accuracy obtained from EDHS-HIV/AIDS data set using wrapper based feature selection with Random Forest, Gradient Boosting, K-Nearest Neighbor, Support Vector Machine, AdaBoost, Naive Bayes and Logistic Regression classifier is **0.889**, **0.888**, **0.868**, **0.832**, **0.796**, **0.767** and **0.746** respectively. The above accuracy score are obtained from the selected feature with 20 features from the original features compared with selected features of Forward Feature Selection (FFS) with feature **15** and **10**. Similarly, in the Backward Feature Selection methods, the experimental result is obtained with same classifiers which are used in FFS methods and the result with Random Forest, Gradient boosting, K-Nearest neighbors, Ada-boost, Support Vector Machine, Logistic Regression and Naive Bayes are **0.907**, **0.895**, **0.875**, **0.756**, **0.748**, **0.701** and **0.701** respectively. This accuracy score is obtained from the selected features of **20** features compared with **15** and **10** selected features from original features of the EDHS-HIV/AIDS data set.

In this Experiments, we also test Recursive feature selection methods with **20**, **15** and **10** features and experiments are done with same classifier as taken from FFS and BFS methods. The experimental result are obtained with Random Forest, Gradient Boosting, K-Nearest neighbors, Ada-boost, Support Vector Machine, Logistic Regression and Naive Bayes is **0.948**, **0.909**, **0.903**, **0.821**, **0.815**, **0.714** and **0.713** respectively.

TABLE 4.1: Result of Wrapper based feature selection methods on EDHS-HIV/AIDS dataset

Classification Algorithms	**Feature Selection Methods**	**No. of Selected Features**	**Evaluation Metrics**				
			Acc	**Pre**	**Rec**	**F1-score**	**ROC**
Random Forest (RF)	**FFS**	**20**	**0.889**	**0.833**	**0.974**	**0.898**	**0.89**
		15	0.870	0.899	0.833	0.864	0.87
		10	0.857	0.898	0.807	0.850	0.86
	BFS	**20**	**0.907**	**0.860**	**0.971**	**0.912**	**0.91**
		15	0.875	0.896	0.849	0.871	0.88
		10	0.800	0.800	0.800	0.800	0.80
	RFS	**20**	**0.948**	**0.918**	**0.984**	**0.950**	**0.95**
		15	0.923	0.887	0.970	0.926	0.92
		10	0.859	0.803	0.951	0.871	0.86
K-Nearest Neighbour (KNN)	**FFS**	**20**	**0.868**	**0.835**	**0.916**	**0.874**	**0.87**
		15	0.852	0.862	0.841	0.851	0.85
		10	0.848	0.910	0.778	0.839	0.85
	BFS	**20**	**0.875**	**0.837**	**0.932**	**0.882**	**0.88**
		15	0.857	0.880	0.830	0.854	0.86
		10	0.761	0.768	0.753	0.760	0.76

Continued on next page

Table 4.1 – Continued from previous page

Classification Algorithms	Feature Selection Methods	No. of Selected Features	Evaluation Metrics				
			Acc	Pre	Rec	F1-score	ROC
		20	**0.903**	**0.872**	**0.944**	**0.907**	**0.90**
	RFS	**15**	0.885	0.848	0.938	0.891	0.89
		10	0.848	0.805	0.919	0.858	0.85
		20	**0.832**	**0.775**	**0.936**	**0.848**	**0.83**
	FFS	**15**	0.830	0.774	0.932	0.846	0.83
Support		**10**	0.830	0.775	0.929	0.845	0.83
Vector Machine		**20**	**0.748**	**0.677**	**0.949**	**0.790**	**0.75**
(SVM)	**BFS**	**15**	0.739	0.675	0.923	0.780	0.74
		10	0.739	0.675	0.920	0.779	0.74
		20	**0.81**	**0.754**	**0.932**	**0.834**	**0.82**
	RFS	**15**	0.826	0.778	0.912	0.840	0.83
		10	0.821	0.778	0.909	0.835	0.82
		20	**0.767**	**0.815**	**0.862**	**0.838**	**0.77**
	FFS	**15**	0.759	0.756	0.770	0.763	0.76
		10	0.756	0.750	0.773	0.761	0.76
Naive Bayes		**20**	**0.701**	**0.682**	**0.752**	**0.715**	**0.70**
(NB)	**BFS**	**15**	0.666	0.628	0.817	0.710	0.67
		10	0.645	0.559	0.877	0.712	0.65
		20	**0.713**	**0.670**	**0.853**	**0.750**	**0.71**
	RFS	**15**	0.714	0.665	0.868	0.753	0.71
		10	0.713	0.665	0.876	0.756	0.71
		20	**0.746**	**0.726**	**0.805**	**0.764**	**0.75**
	FFS	**15**	0.757	0.755	0.765	0.760	0.78
		10	0.755	0.753	0.773	0.763	0.76
Logistic		**20**	**0.701**	**0.684**	**0.746**	**0.714**	**0.70**
Regression	**BFS**	**15**	0.689	0.713	0.642	0.676	0.69
(LR)		**10**	0.661	0.656	0.691	0.673	0.66
		20	**0.714**	**0.671**	**0.851**	**0.751**	**0.71**
	RFS	**15**	0.715	0.667	0.870	0.755	0.71
		10	0.715	0.669	0.876	0.759	0.71
		20	**0.796**	**0.779**	**0.838**	**0.807**	**0.80**
	FFS	**15**	0.794	0.775	0.832	0.802	0.79
		10	0.776	0.796	0.752	0.774	0.78
AbaBoost		**20**	**0.756**	**0.755**	**0.767**	**0.761**	**0.76**
(AB)	**BFS**	**15**	0.736	0.743	0.733	0.738	0.74
		10	0.679	0.671	0.709	0.689	0.68
		20	**0.821**	**0.799**	**0.863**	**0.830**	**0.82**
	RFS	**15**	0.803	0.786	0.839	0.812	0.80
		10	0.776	0.771	0.799	0.785	0.78

Continued on next page

Table 4.1 – Continued from previous page

Classification Algorithms	Feature Selection Methods	No. of Selected Features	Evaluation Metrics				
			Acc	Pre	Rec	F1-score	ROC
Gradient Boosting (GB)	FFS	**20**	**0.888**	**0.912**	**0.962**	**0.962**	**0.89**
		15	0.875	0.911	0.833	0.870	0.86
		10	0.861	0.909	0.807	0.855	0.86
	BFS	**20**	**0.895**	**0.914**	**0.875**	**0.894**	**0.90**
		15	0.877	0.896	0.856	0.875	0.88
		10	0.799	0.792	0.816	0.804	0.80
	RFS	**20**	**0.909**	**0.904**	**0.917**	**0.911**	**0.91**
		15	0.897	0.901	0.894	0.897	0.90
		10	0.876	0.904	0.847	0.875	0.88

According to our experiments, we compare the wrapper methods them selves based on the accuracy to recommend or suggest which methods of wrapper based feature selection is best to select or choose the optimal features from EDHS-HIV/AIDS. As the massive experiments were conducted and the results presented in Table 4.1, we observe RFS is best with compared to FFS and BFS based on the performance evaluation criteria specifically by accuracy.

4.3 Conclusion

In this research, a proposed feature selection approach was conducted using wrapper-based feature selection algorithms to predict the individual status or test outcome of HIV / AIDS from the EDHS data set. The research uses three methods of features selection under the wrapper base to classify the data set namely FFS, BFS and RFS. In our work, We assess the selected features and test their classification performance output using Random Forest, K-Nearest neighbors, Support Vector Machine, Gradient Boosting, AdaBoost, Naive Bayes and Logistic Regression classifier. The performance was measured by five evaluation metrics namely: Accuracy, Precision, Recall, F1-measure and ROC.

From the experiment, Random Forest, K-Nearest neighbors and Gradient Boosting classifiers have higher accuracy levels on EDHS-HIV/AIDS data set than the others after the applying the wrapper based feature selection methods. This study shows that methods of selecting features are capable of improving the learning algorithms efficiency.

Finally, the output of this study can make significant contributions in the prediction of the status of HIV/AIDS result of individuals in health domain research and provide wrapper based feature selection methods for machine learning studies.

Chapter 5

A Comparative Study of Filter and Wrapper Methods on EDHS – HIV/AIDS Dataset

5.1 Introduction

The aim of this research work is present a comprehensive empirical evaluation and comparison of two methods of feature selection i.e. wrapper and filter with the original feature set of the EDHS - HIV/AIDS data set and compare different machine learning algorithms including LR, KNN, SVM, DT, RF and GB for prediction on EDHS data set. These Machine Learning algorithms were implemented for prediction of HIV/AIDS status.

5.2 Proposed Methodology

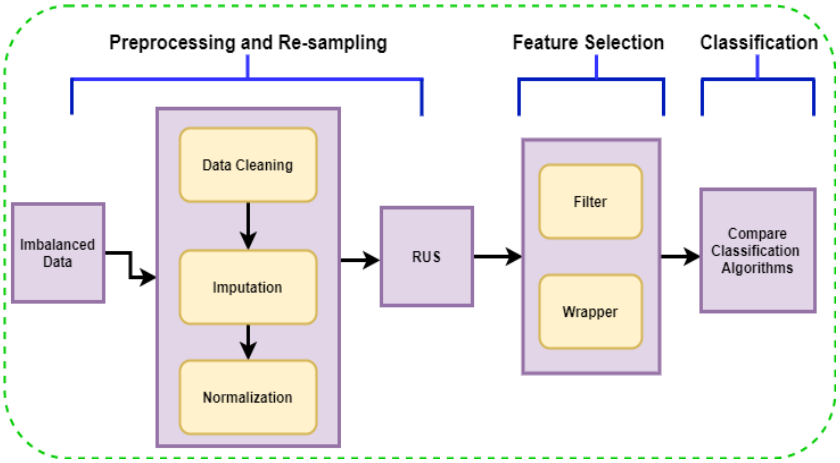

FIGURE 5.1: The proposed methodology

5.2.1 Preprocessing

Re-sampling: Random under sampling (RUS) re-sampling methods have been extended on both the decreased data and the initial (unbalanced) data sets.

As the classification are not fairly represent (i.e. **55,209:23,668** negative – positive instances), we under sampled the majority class using random under-sampling technique (RUS). As a result, we obtained a balanced dataset containing **38,563:38,563** negative – positive instances. The distribution in under random sampling technique is using **0.5** ratios. Table 5.1. Shows the used data set for this proposed methods.

TABLE 5.1: Characteristics of the EDHS-HIV/AIDS Dataset

Dataset	Instances	Features	Training Set	Test Set	Negative	Positive
Full Dataset	77,877	26	63,101	15,776	55,209	23,668
Filter Set	77,126	21	61,700	15,426	38,563	38,563
Wrapper Set	77,126	21	61,700	15,426	38,563	38,563

5.3 Experimental Results and Performance Evaluation

For each classifier, the feature selection evaluates reflecting the Wrapper and Filter approaches were used to define the minimum features that resulted in the maximum Receiver Operating Characteristic (ROC) and the accuracy of the classifier was computed. As described below, the technique varies marginally between the selections methods for the feature. Accuracy and ROC are standard output assessment metrics extracted from a confusion matrix [52].

For the wrapper method, we used a Backward Feature subset selection criterion on the full training dataset **(77,126 x 21)** with best selected features. For filter methods we used univariate feature selection methods which are ranked by specific criteria. In filter, we used chi-square methods to select the top N features from the full training dataset **(77,126 x 21)**.

The analysis is initiated in filter method by rating the function in descending order according to the information gain of the class. Thereafter, a backward search was implemented in the wrapper process to define the minimum features that resulted in the maximum ROC for each classifier. Basically, the features were removed sequentially, and the precision of the classifier was determined. It began with the feature exclusion with the lowest ROC, and only ended until the strongest ROC parameter was used. For each classifier, this procedure was repeated twenty times (i.e., the cumulative number of features in the methods of choosing backward tasks).

In the next section we show the results of full feature set, filter set and wrapper set accordingly.

5.3.1 Results from full feature set

The results obtained from the original data set with six classification algorithms of ROC are shown in Figure 5.2 and the results of each classifier are showed in Table III. As the result of these original data set presented in this section are the base of comparing the two methods. Figure 5.2 and Table 5.2 shows the experiment result of the full training set from EDHS-HIV/AIDS data set with original feature set. As we explained in section 5.3, we took the results of the accuracy and ROC as a metrics.

Accordingly, the results obtained from HIV/AIDS data set without applying feature selection methods (i.e original feature set) with GB, SVM, DT, RF, KNN and LR classifiers are **0.86**, **0.84**, **0.82**, **0.79**, **0.79** and **0.72** respectively with accuracy. And in the ROC the result of SVM, GB, LR, DT, RF and KNN classifiers are **0.9151**, **0.8651**, **0.8360**, **0.8183**, **0.8036** and **0.7935** respectively. The above scores are obtained from the original feature set.

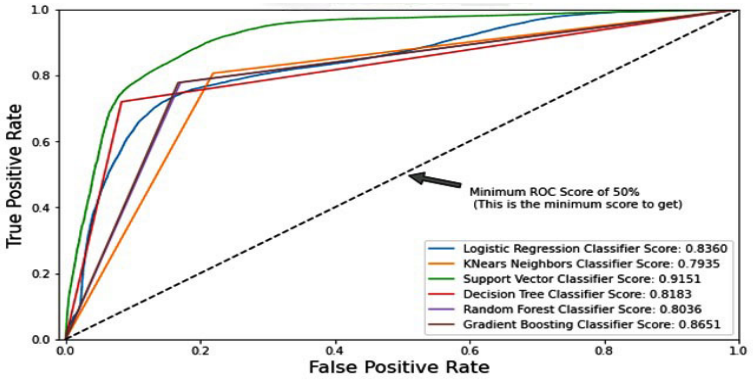
True Positive Rate
False Positive Rate
Minimum ROC Score of 50%
(This is the minimum score to get)
Logistic Regression Classifier Score: 0.8360
KNears Neighbors Classifier Score: 0.7935
Support Vector Classifier Score: 0.9151
Decision Tree Classifier Score: 0.8183
Random Forest Classifier Score: 0.8036
Gradient Boosting Classifier Score: 0.8651

FIGURE 5.2: The ROC curve on full training set (feature set)

TABLE 5.2: Results of both classifier on full training set with original feature set

	Classifiers					
Metrics	**LR**	**KNN**	**SVM**	**DT**	**RF**	**GB**
Accuracy	**0.72**	**0.79**	**0.84**	**0.82**	**0.79**	**0.86**
Precision	0.73	0.79	0.85	0.83	0.79	0.87
Recall	0.72	0.79	0.84	0.82	0.79	0.86
F1-Score	0.72	0.79	0.84	0.82	0.79	0.86
ROC	**0.8360**	**0.7935**	**0.9151**	**0.8183**	**0.8036**	**0.8651**

5.3.2 Results from Features Selection

This section presented the results of the feature selection method applying in this study. The methods are filter and wrapper.

5.3.2.1 Wrapper

Figure 5.3 and Table 5.3 shows the experiment result of the wrapper set from EDHS-HIV/AIDS data set. As we explained in section **??**, we had taken the results of the accuracy and ROC as a metrics. The results obtained from HIV/AIDS data set with applying Wrapper feature selection methods with SVM, GB, KNN, DT, RF and LR classifiers are **0.86**, **0.86**, **0.83**, **0.83**, **0.79** and **0.76** respectively with accuracy. And in the ROC the result of SVM, GB, KNN, LR, DT and RF are **0.9222**, **0.8644**, **0.8386**, **0.8365**, **0.8264** and **0.8036** respectively. The above scores are obtained from the selected feature with **20** features from the original feature set compared with selected features of wrapper methods. In our case we used backward feature selection criterion to select the best feature.

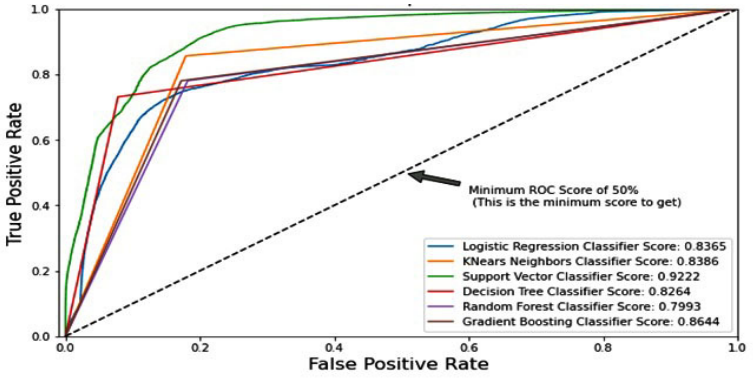
True Positive Rate
False Positive Rate
Minimum ROC Score of 50%
(This is the minimum score to get)
Logistic Regression Classifier Score: 0.8365
KNears Neighbors Classifier Score: 0.8386
Support Vector Classifier Score: 0.9222
Decision Tree Classifier Score: 0.8264
Random Forest Classifier Score: 0.7993
Gradient Boosting Classifier Score: 0.8644

FIGURE 5.3: The ROC curve on wrapper methods

TABLE 5.3: Result of wrapper feature selection on both classifiers.

	Classifiers					
Metrics	**LR**	**KNN**	**SVM**	**DT**	**RF**	**GB**
Accuracy	**0.76**	**0.83**	**0.86**	**0.83**	**0.79**	**0.86**
Precision	0.76	0.83	0.86	0.84	0.79	0.86
Recall	0.76	0.83	0.86	0.83	0.79	0.86
F1-Score	0.76	0.83	0.86	0.83	0.79	0.86
ROC	**0.8365**	**0.8386**	**0.9222**	**0.8264**	**0.8036**	**0.8644**

5.3.2.2 Filter

Figure 5.4 and Table 5.4 shows the experiment result of the wrapper set from EDHS-HIV/AIDS data set. As we explained in section 5.3, we had taken the results of the accuracy and ROC as a metrics. The results obtained from HIV/AIDS data set with applying Wrapper feature selection methods with GB, SVM, KNN, DT, RF and LR classifiers are **0.86**, **0.85**, **0.83**, **0.82**, **0.80** and **0.75** respectively. Where as in the ROC the result of SVM, GB, LR, KNN, DT and RF are **0.9238**, **0.8643**, **0.8331**, **0.8317**, **0.8182** and **0.7997** respectively. The above scores are obtained from the selected feature with 20 features from the original feature set compared with selected features of filter methods. In our case we used univariate feature selection criterion to select the best feature.

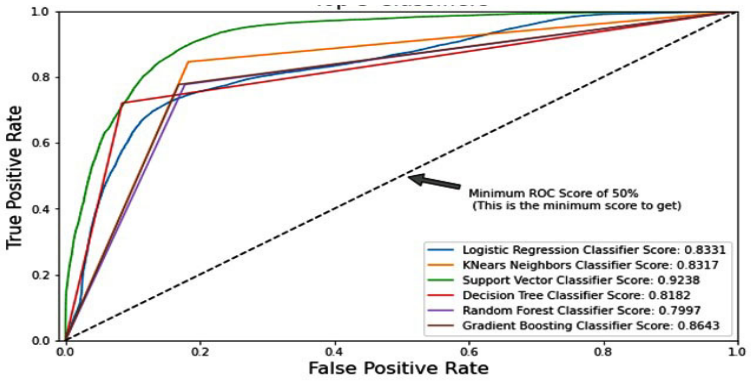
True Positive Rate
False Positive Rate
Minimum ROC Score of 50%
(This is the minimum score to get)
Logistic Regression Classifier Score: 0.8331
KNears Neighbors Classifier Score: 0.8317
Support Vector Classifier Score: 0.9238
Decision Tree Classifier Score: 0.8182
Random Forest Classifier Score: 0.7997
Gradient Boosting Classifier Score: 0.8643

FIGURE 5.4: ROC curve on filter methods

TABLE 5.4: Result of filter feature selection on both classifiers.

	Classifiers					
Metrics	**LR**	**KNN**	**SVM**	**DT**	**RF**	**GB**
Accuracy	**0.75**	**0.83**	**0.85**	**0.82**	**0.80**	**0.86**
Precision	0.75	0.83	0.86	0.83	0.80	0.86
Recall	0.75	0.83	0.85	0.82	0.80	0.86
F1-Score	0.75	0.83	0.85	0.81	0.80	0.86
ROC	**0.8331**	**0.8317**	**0.9238**	**0.8182**	**0.7997**	**0.8643**

5.3.3 Performance Evaluation

The result of classification using the best features sets drawn from both feature selection methods were compared against those obtained on full dimensional training set. The comparison is taken using the Accuracy score and ROC curve without including the other performance metrics discuss above. As the result shows that in Table VI, there is a difference among the two methods. In wrapper methods the classifiers such as LR, SVM, DT and RF are **0.76**, **0.86**, **0.83** and **0.81** respectively. It scored better Accuracy than filter, whereas KNN and GB have same results in both methods. In ROC metrics, the wrapper scored better than Filter. The result in LR, KNN, DT

and GB is **0.8365**, **0.8386**, **0.8264** and **0.8644** respectively, but with SVM and RF classifier filter method scored better i.e. **0.9235** and **0.7997** respectively. Results revealed that wrapper selected subset performed slightly better than filter subset with accuracy. Both methods improved the initial model on full training set by a small margin. Performance results full data set, filter set and wrapper set is shown in Table 5.5.

TABLE 5.5: Performance result of full data set, filter set and wrapper set on both classifiers

Dataset	Classifier	Metrics	
		Accuracy	ROC
Full Dataset	**LR**	0.72	0.8360
	KNN	0.79	0.7935
	SVM	0.84	0.9151
	DT	0.72	0.8183
	RF	0.79	0.8036
	GB	0.86	0.8651
Filter set	**LR**	0.75	0.8331
	KNN	0.83	0.8317
	SVM	0.85	0.9238
	DT	0.82	0.8182
	RF	0.80	0.7997
	GB	0.86	0.8643
Wrapper set	**LR**	0.76	0.8365
	KNN	0.83	0.8386
	SVM	0.86	0.9222
	DT	0.83	0.8264
	RF	0.81	0.7993
	GB	0.86	0.8644

5.4 Conclusion

In this research work, we presented the empirical comparison of the feature selection methods between filter and wrapper for imbalanced data set of EDHS data sets. Like with controlled learning methods, selection of features will improve the consistency of the outputs while reducing the learning process difficulty. Wrapper methods use the induction algorithm bias to pick features and generally work better, particularly on datasets where algorithm-specific learning of optimal feature sets. Filter methods can do well on real-world datasets, as they can also classify strong sets of features. But the computing cost of wrapper approaches on massive data sets is prohibitive.

This paper presents an algorithm that uses six different classification algorithms namely: LR, KNN, DT, GB, SVM and RF to compare the two types of feature section methods in EDHS data sets. As the result shows that in Table VI, there are some differences among the two methods. In wrapper, the classifiers such as LR, SVM, DT and RF are scored better Accuracy than Filter, whereas KNN and GB have same results in both methods. In ROC metrics, the wrapper scored better than Filter. The result in LR, KNN, DT and GB is slightly different. But with SVM and RF filter scored better.

In summary, wrapper approaches work pretty well despite the minimal knowledge they utilize and their substantially reduced difficulty with regard to the approaches of filtering is valuable in practice. The filter has the ability to allow an effective selection but output or even chosen features that vary between classifiers even while using the same training sets. The difference in the actions of both approaches indicates that a common definition for choosing useful features through multiple learning algorithms cannot occur.

Chapter 6

Performance Evaluation of Classification Models for HIV/AIDS Dataset

6.1 Introduction

This paper aims at using machine learning algorithms to classify the person was being tested for HIV (i.e. either negative or positive) using HIV/AIDS data and the classification power of the various models based on the machine learning algorithm is compared.

6.2 Proposed Methods

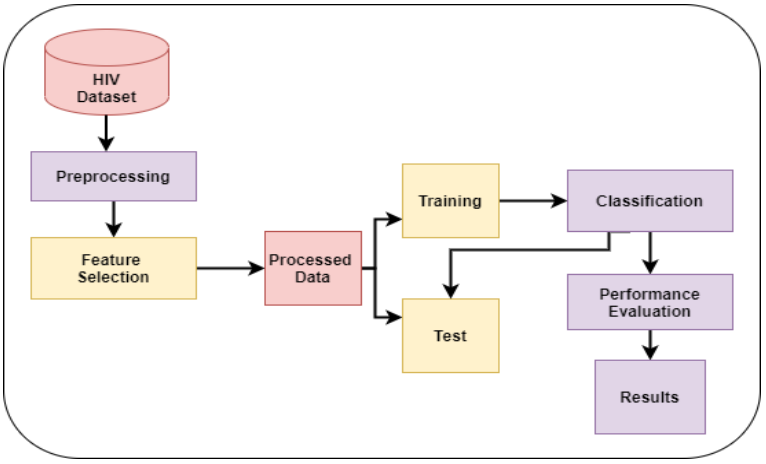

FIGURE 6.1: Proposed Methodology

6.3 Experimental Result

As Table 6.1 shows that, the performance of all classifiers has better results for all performance metrics. The proposed approach efficiency for the accuracy of the training value is higher than **0.80%** is achieved except the Naive Bayes. The other performance metrics like recall, F1-score, and AUC were also commendable. As

shown in Table 6.1, the results of the Random Forest, Extra Tree and Decision Tree model is best with a training accuracy of **0.9755** and the score of Bagging, K-nearest neighbor, Gradient Boosting, Logistic Regression, SVM, Ada Boost, and Naive Bayes are **0.9669**, **0.8822**, **0.8774**, **0.8285**, **0.8234**, **0.8086** and **0.7731** respectively. In addition, the testing accuracy Gradient Boosting, Random Forest, Extra Tree, Bagging, K-nearest neighbor, Logistic Regression, Decision Tree, SVM, Ada Boost and Naive Bayes are scored **0.8721**, **0.8657**, **0.8590**, **0.8530**, **0.8297**, **0.8289**, **0.8256**, **0.8226**, **0.8110** and **0.7784** respectively.

TABLE 6.1: Results of Classification algorithm based on different performance metrics

Model	Training Accu	Testing Accu	Precision	Recall	F1-score	AUC
Gradient Boosting **(GB)**	0.8774	**0.8721**	0.836567	0.712047	0.769301	0.826303
Random Forest **(RF)**	0.9755	**0.8657**	0.783413	0.762016	0.772566	0.835997
Extra Tree **(ET)**	0.9755	**0.8590**	0.774024	0.746983	0.760263	0.826898
Bagging	0.9669	**0.8530**	0.769024	0.727504	0.747688	0.817068
K-nearest Neighbor **(KNN)**	0.8822	**0.8297**	0.751917	0.643659	0.693589	0.776457
Logistic Regression **(LR)**	0.8285	**0.8289**	0.785231	0.589879	0.673679	0.7604691
Decision Tree **(DT)**	0.9755	**0.8256**	0.716228	0.691510	0.703652	0.787219
Support Vector Machine **(SVM)**	0.8234	**0.8226**	0.786544	0.559390	0.653799	0.747260
AdaBoost **(AB)**	0.8086	**0.8110**	0.771607	0.523608	0.623865	0.728691
Naive Bayes **(NB)**	0.7731	**0.7784**	0.650184	0.562354	0.603088	0.716534

Figure 6.2 contrasts the performance results of the ten ML techniques using AUC/ROC. AUC/ROC curves allow the depiction of the relationship of the TPR and FPR, whereas AUC helps to evaluate different algorithms or combinations of hyper-parameters. As shown in Figure 6.2, except Naive Bayes the other models perform well; and, the AUC/ROC curves make very little difference. Nevertheless, there is a relatively higher positive curve on the ROC curve compared to the Random Forest model, suggesting that the variant provides a better true positive rate. It may also be concluded, in combination with the obtained results for the performance metrics, that the Random Forest model can classify HIV tests more effectively than the other models.

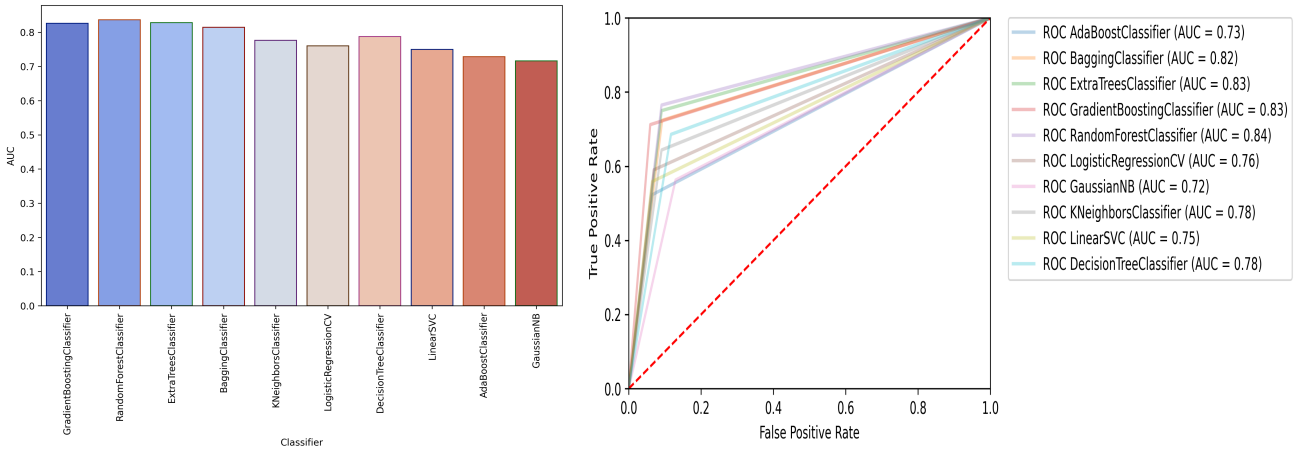

FIGURE 6.2: Comparison of ten classifier with AUC/ROC. Area under curve (left), ROC curve (right)

On the other hand, we analyzed the performance of the method by the use of precision, recall, and f1-score using cross-validation. The experiment of the work also shows the performance or result from a comparison of all ten machine learning algorithms with precision, recall, and f1-score. Depending on each classifier algorithm the result of the performance metrics for the above three is different. Figure 6.3 and Figure 6.4 shows the results of

each classifier with its precision, recall, and f1-score respectively. In precision, the Gradient Boosting classifier is best compared to the rest and in recall and f1-Score, Random Forest is best.

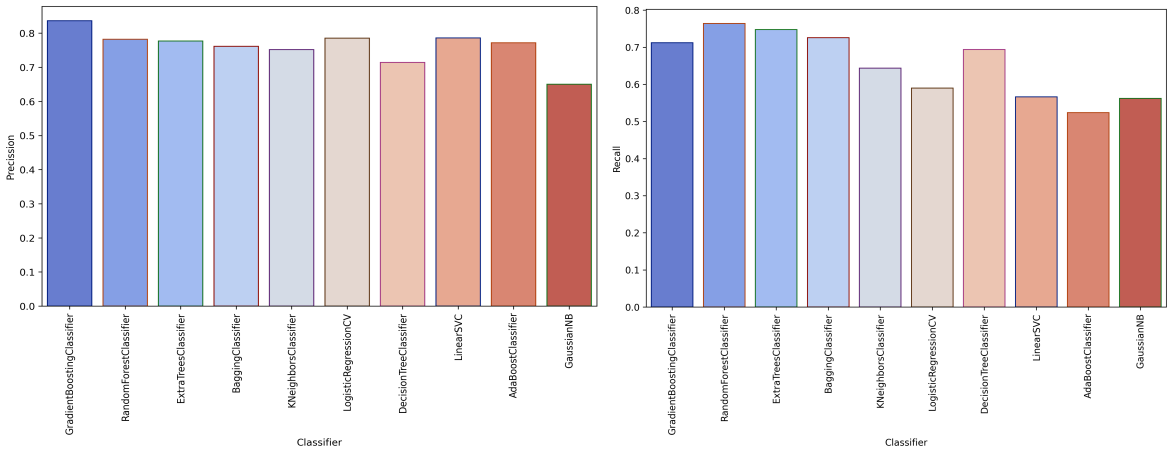

FIGURE 6.3: Comparison of ten classifier with Precision and recall. Precision (left), Recall (right)

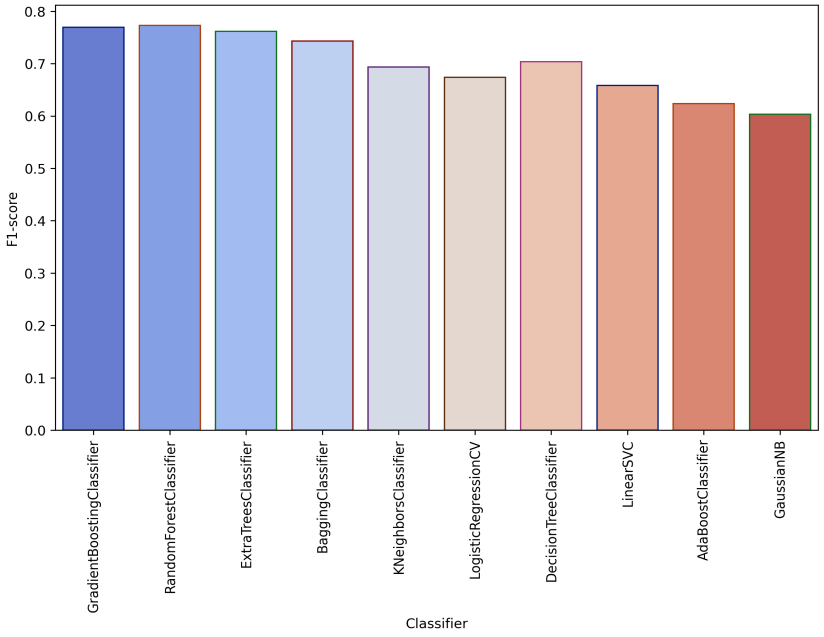

FIGURE 6.4: Results of each classifier F1-score.

Finally, in this work, binary class data are classified using ten different machine learning algorithms. The classification results obtained from all classifiers are promising; (above 80%) as shown in Table 6.1. To validate our work, we employed a confusion matrix that indicates correct and wrong prediction samples in the test data. Figure 6.5, shows the confusion matrix of HIV/AIDS data.

6.3.1 Comparison with previous work

This section discuses the performances of our proposed method and comparison with the state-of-the-art methods. The comparison is done through the method which is applied in each work. We were doing a careful investigation of each paper in terms of methods, dataset, algorithms used, and the tools for experiments and compare our proposed methods of classification problem with the related work based on the accuracy result. In our research, we applied a feature selection methods to improve the performances of the classifiers. An optimal

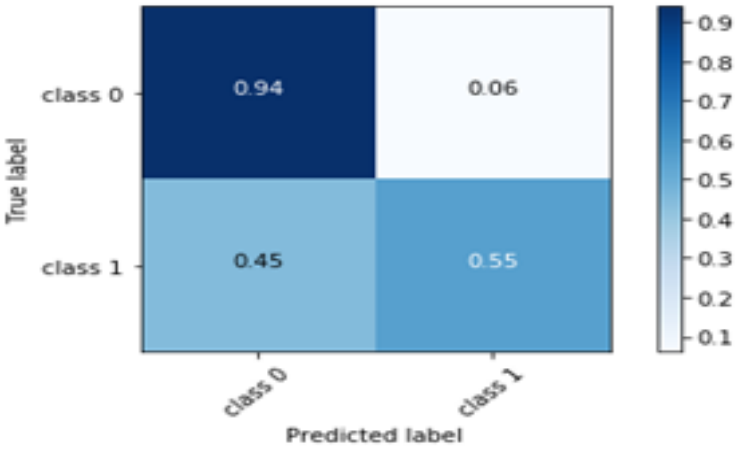

FIGURE 6.5: Confusion matrix the HIV/AIDS dataset.

performance of our method is achieved when we used feature selection to get optimal features. In the investigation the proposed method classifies are better with their accuracy. This shows that the proposed method is performed better than other of the recent related works.

Table 6.2 presents a comparative analysis of our method with some selective and related works with respect to accuracy.

6.4 Conclusion

This study presents a proposed methodology to classify the status of HIV/AIDS using different machine learning algorithms. The considered machine learning models used in this study is: Random Forest, Extra Tree, Decision Tree, Bagging, K-nearest neighbor, Gradient Boosting, Logistic Regression, SVM, AdaBoost, and Naive Bayes. The purpose of this work was to compare algorithms with different performance metrics. The performance metrics such as accuracy, recall, precision, f1-score, AUC/ROC, and confusion matrix were employed to confirm the validity of the method. The experimental results of the proposed method exhibit better classification of the given datasets i.e. HIV/AIDS, by scoring training accuracy up to 0.9755, and in testing accuracy, it scored up to **0.8721**. In the comparison of the-state-of-the-art our method score better results in terms of accuracy. The study can be considered as a pilot run and its approach may prove to be beneficial in further studies.

TABLE 6.2: Comparison of the proposed method with selected related research works with respect to accuracy

Author	Method	Dataset	Tool	Algorithms	Accuracy
Hailu Tsegaye [35]	CRISP-DM	EDHS, 2011	WEKA	Induction DT	0.79
				NN	0.78
				LR	0.74
				Random Tree DT	0.96
Eliyas Lemuye [39]	CRISP-DM	VCT	WEKA	J48	0.82
				ID3	0.78
Teklu Urgesa [37]	CRISP-DM	Adama and Assela Hospital ART	WEKA	Binary DT (Pruned)	0.76
				Binary DT (Un-pruned)	0.69
				General DT (Pruned)	0.78
				General DT (Un-pruned)	0.70
Leke B. et al [27]	RBF	SA Antinatal 2011	WEKA	MLP	0.84
Koccak. et al. [23]	Statistical Method	UCI	-	FIN-CLF	0.82
				DT	0.80
				KNN	0.79
				NB	0.69
Birhanu Fetene [38]	CRISP-DM	VCT	WEKA	J48	0.78
				NB	0.78
				MLP	0.76
Annang, D. A. [34]	CRISP-DM	VCT	WEKA	J48	0.83
				RT	0.78
				NB	0.79
				NN	0.77
				LR	0.81
Bao, Y., et al [40]	Machine Learning	MSM	-	GBM	0.76
				DL	0.75
				RF	0.72
				MLR	0.69
				XGB	0.71
Proposed Method	Machine Learning	EDHS 2000, 2005, 2011 and 2016	Python	GB	**0.872**
				RF	**0.865**
				ET	**0.859**
				Bagging	**0.853**
				KNN	**0.829**
				LR	**0.828**
				DT	**0.825**
				SVC	**0.822**
				AB	**0.811**
				NB	**0.778**

Chapter 7

Classification of Binary Class HIV/AIDS Test Results Using Ensemble Learning Models

7.1 Introduction

This research work aims to present an ensemble learning method for binary data classification on the HIV dataset and compares the results of the ensemble classifier before and after feature selection was applied. The research uses five established classifiers, namely Gradient Boosting (GB), k-Nearest Neighbor (KNN), Random Forest (RF), Extra Tree (ET), and Multi layer Perceptron (MLP).

7.2 Proposed Methodology

In this study, the ensemble learning approach is proposed for the classification of binary class data on the basis of five existing classifiers. In the case of the classification query, the classifier is usually given a known training dataset for training and an unknown dataset to validate the model. The 10-fold cross-validation approach was used to train and validate the classifier. Cross-validation describes a dataset to assess the model during the training process to solve the issue of over-fitting. The process for the proposed model is shown in Figure 7.1.

7.3 Experimental Results and Discussion

Experimental findings of the proposed model on EDHS-HIV/AIDS datasets are discussed in this section. The findings obtained by the proposed methods are presented below. The findings for each classifier are presented in Table 7.1 before and after the selection of the feature. The 10-fold cross-validation process is used to divide the data into training and test set.

In the experiment, the result shows that all classifiers can achieve better results by applying the backward feature selection method with all performance metrics. Tables 7.1 illustrate the classification results of the established classifiers before and after applying feature selection methods. Besides, Figures 7.2 and Figure 7.3 illustrate the comparison results of all models before and after feature selection is applied to the ROC/AUC and precision/recall respectively.

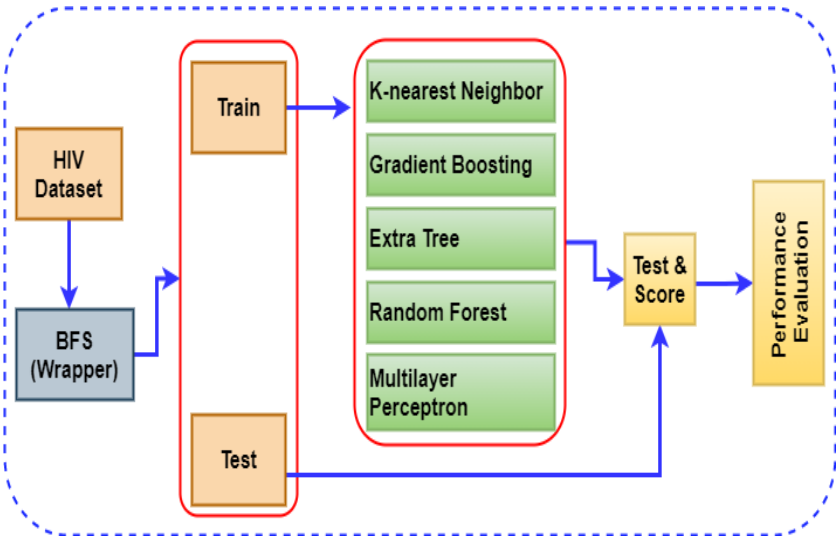

FIGURE 7.1: Proposed methodology of ensemble learning method

TABLE 7.1: Classification report of all models before and after feature selection

Model	Results Before Feature Selection					Results After Feature Selection				
	Accu	**Prec**	**Rec**	**F1-sco**	**AUC**	**Accu**	**Prec**	**Rec**	**F1-sco**	**AUC**
Gradient Boosting **(GB)**	**0.81**	0.82	0.71	0.79	0.85	**0.88**	0.87	0.83	0.87	0.94
Multilayer Perceptron **(MLP)**	**0.83**	0.81	0.75	0.82	0.87	**0.87**	0.85	0.83	0.87	0.94
Random Forest **(RF)**	**0.83**	0.81	0.76	0.82	0.87	**0.86**	0.84	0.84	0.86	0.92
Extra Tree **(ET)**	**0.82**	0.81	0.75	0.82	0.87	**0.86**	0.83	0.83	0.86	0.90
K-Nearest Neighbour **(KNN)**	**0.80**	0.77	0.75	0.80	0.82	**0.83**	0.80	0.78	0.82	0.87

The classification report before applying the feature selection method for each accuracy level, the MLP and RF classifiers among other classifiers can achieve comparatively better accuracy. Nevertheless, in all metrics, the results are improved from each model. From the result of the experiment GB is improved by **0.07%**, MLP by **0.04%**, RF by **0.03%**, ET by **0.04%**, and KNN by **0.03%**. The results achieved the highest accuracy for all classifications following the implementation of the feature selection method (i.e. the backward criterion of the wrapper approach). Furthermore, accuracy alone is not considered to be a reasonable indicator of the outcome of the classifiers, so other performance metrics, such as precision, recall, and f1-score have been considered.

When tested using the receiver operating characteristic curve (ROC), the efficiency of the proposed model registers enhances the outcome. In addition to other performance metrics, the ROC curve is used and has a high degree of tolerance when classifying low-class imbalances results. ROC curve for five of the GB, MLP, KNN, ET, and RF classifiers on the binary class HIV / AIDS dataset after the feature selection is shown in Figure 7.2. Table 7.1 shows the percentage output of all classifiers. Based on the results obtained, it is proposed that the use of the ensemble classification method would increase the performance of the prediction, which is highly anticipated.

On the other hand on the model evaluation process, we evaluate each model by learning curve. The learning curve illustrates the predictive accuracy of the test examples as a function of the number of training. Figures 7.4, 7.5, 7.6, 7.7, and 7.8 shows the learning curve of all models.

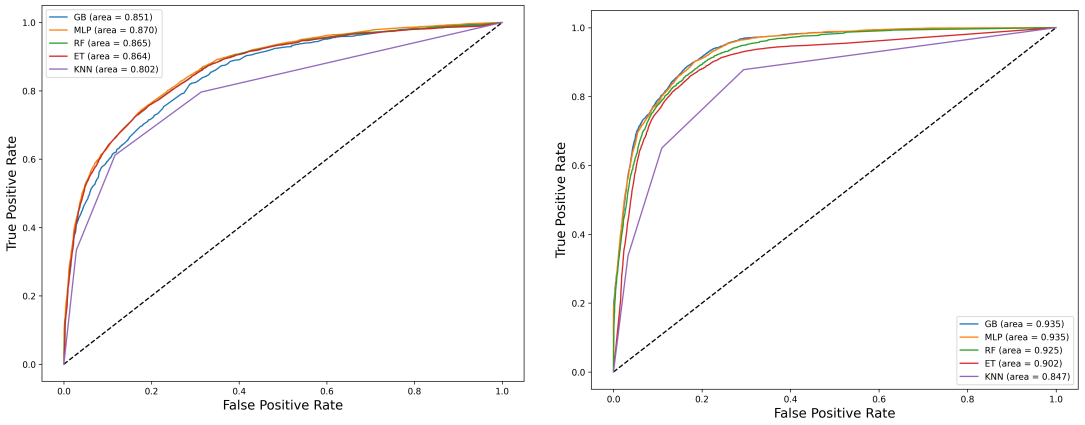

FIGURE 7.2: ROC/AUC results of the five models before and after feature selection: before (left) , after (left) .

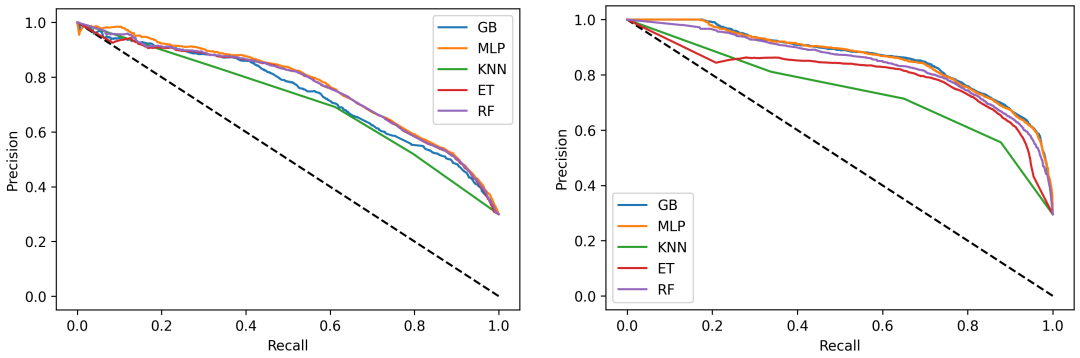

FIGURE 7.3: Precision and recall results of the five models before and after feature selection: before (left), after (right)

The last stage of this experiment is the model scoring, which is based on the actual and predicted target column values in the test set. The whole model labeling method consists of a match count which is how many data rows are properly classified, and how many data rows are misclassified by the model. Figure 7.9 shows the confusion matrix using five established models on the HIV/AIDS dataset of positive and negative class. After the feature selection technique is applying, the number of correctly classified samples is increased.

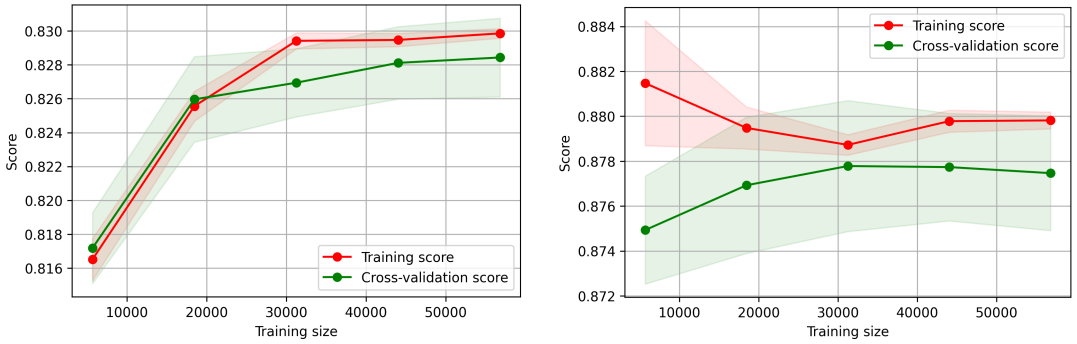

FIGURE 7.4: The learning curve of Gradient Boosting (GB) before and after feature selection: before (left), after (right).

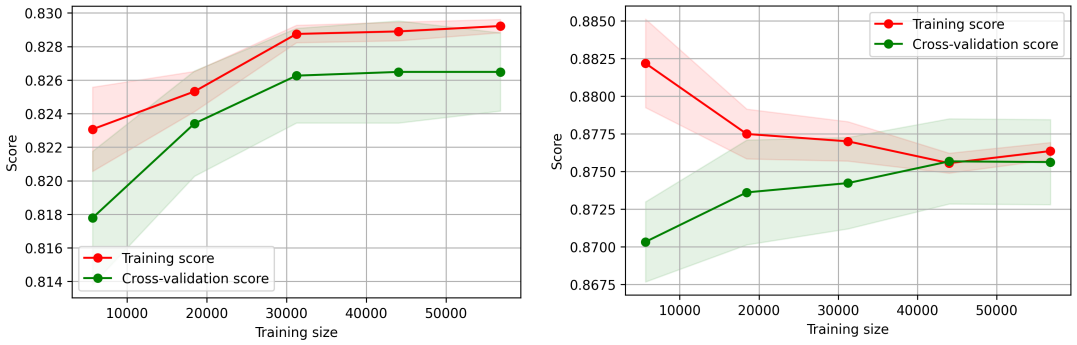

FIGURE 7.5: The learning curve of Multilayer Perceptron (MLP) before and after feature selection: before (left), after (right).

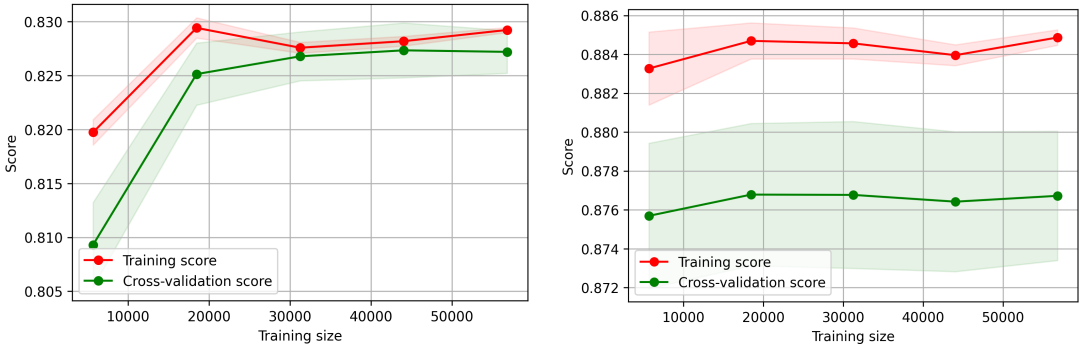

FIGURE 7.6: The learning curve of Random Forest (RF) before and after feature selection: before (left), after (right).

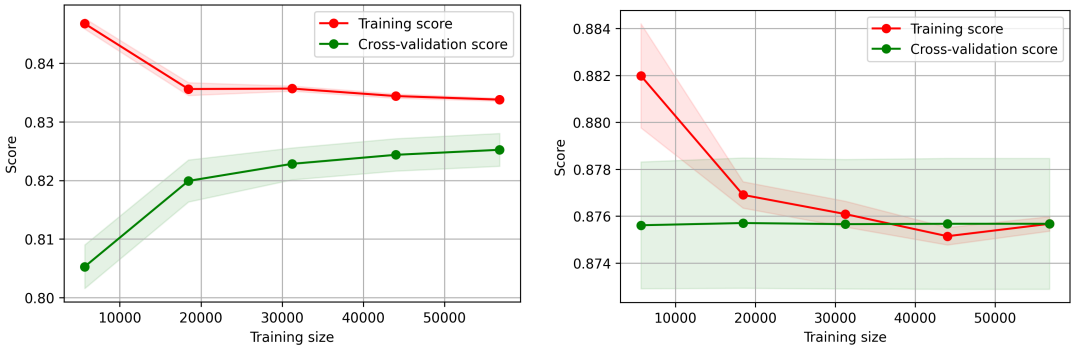

FIGURE 7.7: The learning curve of Extra Tree (ET) before and after feature selection: before (left), after (right).

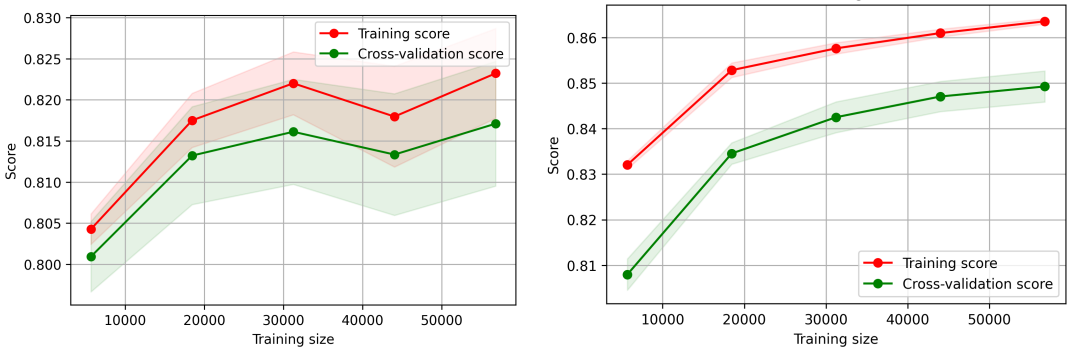

FIGURE 7.8: The learning curve of K-nearest neighbor (KNN) before and after feature selection: before (left), after (right).

FIGURE 7.9: Confusion Matrix for five established classifiers after feature selection applied.

7.4 Conclusion

In this article, an ensemble learning approach based on the binary class data classification of the HIV/AIDS test status is presented. The Ensemble Learning Approach allows more than one classifier at a time in one or more datasets to maximize the output of a single classifier. To address the issue of model over fitting, a selection of features is made using the Backward Feature Selection (BFS). The 10-fold cross-validation procedure was used to train and verify the algorithm that checks the algorithm 9 times and in the 10^{th} iteration, the evaluation is performed. At least each sample is tested. The study uses EDHS-HIV/AIDS dataset for model training and testing. From the analysis, it is noted that both cross-validation and the features selection contribute a great deal to improving the performance measures of the mentioned classifiers. From the point of view, we conclude that a comprehensive and reliable model needs to be developed to reduce minority class's misclassification. Metrics such as accuracy, recall, precision, f1-score, and ROC/AUC are used to evaluate the model performances. The confusion matrix is also used to display properly classified and misclassified samples in both pre- and post-selection of the feature.

Chapter 8

Grid Search In Hyperparameter Optimization of Machine Learning Models for Prediction of HIV/AIDS Test Results

8.1 Introduction

A Machine Learning (ML) algorithm hyper-parameters tuning is a main feature of the model learning process which known to be best practice for achieving a successful implementation [53].

The main objective of this paper is generating the optimal parameter from the default parameter of each model which are used in the study using grid search methods and applying the optimal parameters to each models to test the effects of our methods for the prediction HIV test results in EDHS dataset.

Our contribution in this work is:

- We collect the raw data from EDHS database and make it suitable for the prediction of HIV test results.
- We propose hyper-parameter tuning with grid search for prediction of HIV status with eight established models.
- We finding the best optimal parameters which is fitted to the best models for HIV test Prediction.
- We conduct extensive experiments to evaluate the performance of our method and comparing our method results to statistical methods.
- We address the HPO's challenges and research directions for research domain.

8.2 Proposed Methodology

This section introduced the proposed methodology for HIV test result prediction. The flow diagram of our method is shown in the Figure 8.1. The considered HIV/AIDS data go through many pre-processing phases and selection of features is performed. It is provided as input to the machine learning models after selecting the appropriate features. Finally, the optimization of parameters is carried out by the grid search to advance the efficiency of the model. The full description of the proposed model is given in this section.

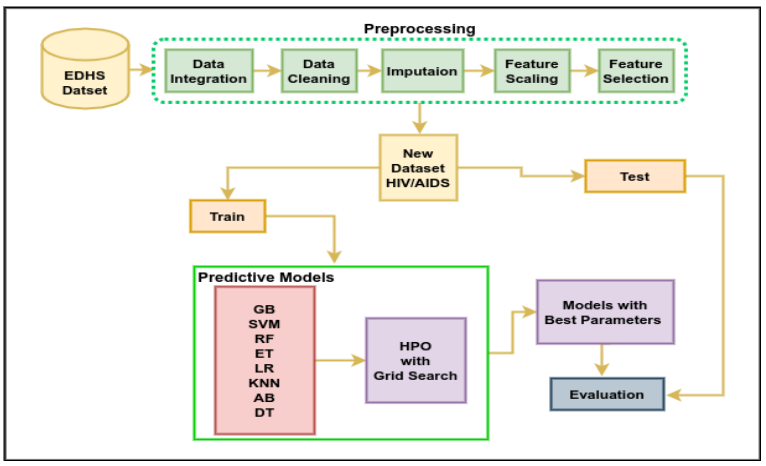

FIGURE 8.1: Flow diagram of the proposed grid search hyper-parameter optimization (GSHPO) method

8.3 Experiments

8.3.1 Experimental Setup

First, the considered HIV/AIDS dataset were provided [54]. Then, the ML algorithms with their objective functions need to be configured. LR, RF, SVM, KNN, DT, GB, AB, and ET are the models to be optimized. In the next phase, evaluation metric and technique are setup. So basically, to evaluate a model's performance, we need some data (input) for which we know the ground truth (label). For this problem, we don't know the ground truth for the test set but we do know for the train set. So the idea is to train and evaluate the model performance on HIV/AIDS data. One thing we do is to split the train set in to two groups, in our case we use **80:20** ratio, the ratio is done by randomly. That means we would train the model on **80%** of a training data and we reserve the rest **20%** for evaluating the model since we know the ground truth for this **20%** data. Then we compare our model prediction with this ground truth (for **20%** data). That's we observe how our model would perform on unseen data. This is the first model evaluation technique. This process is used by sklearn library in a train test split method [55].

We have **78,877** data observations or instances. If k = 10 and each fold would have **78,877/10 = 7,887.7** observations. So basically k-fold cross validation uses fold-1 (**7,887.7** samples) as the testing set and k-1 (9 folds) as the training sets and calculates test accuracy. The process is iterated k times (if k = 10, then 10 times); A different range of observations is considered each time as a validation or test set. This approach results in k test accuracy predictions, which are then averaged out. Table 8.1 illustrates the ten-fold CV score for all models in the study before hyperparameter optimization is applying.

TABLE 8.1: The 10-fold cross validation score for all models before optimized

	1	2	3	4	5	6	7	8
MODEL	GB	SVM	RF	ET	LR	KNN	DT	AB
CV-Score	87.62	87.52	86.72	86.15	82.81	82.65	82.44	80.70

As Table 8.1, the models like GB, SVM, and RF have the highest cross validation score among the models, followed by ET, LR, KNN, DT, and AB. This result shows how our models are score before optimization technique is applied. After that, certain constraints should be met in order to compare various optimization algorithms fairly. Firstly, using the same hyperparameter configuration space, we compare various HPO methods. Table 8.2 shows the basics of the configuration space for ML models. Domain knowledge and manual testing determine

the hyperparameters chosen and their search space [56]. GridSearchCV were considering the values for each hyperparameter that we explicitly define to include the grid search. To get best parameter exhaustive grid-search is processed because it trying every combination of the parameter. Table 8.2 also summarizes the hyperparameter forms of each ML algorithm.

TABLE 8.2: The machine learning models hyperparameter configuration space

Model	Hyperparameter-Search Space	Best Parameter
LR	"penalty":["l1","l2"], "C":np.logspace(0,4,10), "max_iter":[1000]	“C”: [166.8100537203], “max_iter”: [1000], “penalty”: [l2]
GB	"learning_rate":[0.01,0.02], "max_depth":[4,6], "max_features":[1.0,0.3], "min_samp_split":[2,4], "random_state":[42]	“learning_rate”: [0.02], “max_depth”: [6], “max_features”: [1.0], “min_samples_split”:[2], “random_state”: [42]
SVM	"C":np.arange (6,13), "kernel":["linear","rbf"], "gamma":[0.5,0.1,0.001,0.0001]	"C":np.arange (6), "kernel":["rbf"], "gamma":[0.5]
DT	"max_features":["auto", "sqrt","log2"], "min_samp_split":np.arange(2,16), "min_samp_leaf":np.arange(1,12), "gamma":["random_state":[42]	“max_features”: ’auto’, “min_samp_leaf”: 10, “min_samp_split”: 2, “random_state”: 42
KNN	"n_neighbors":np.arange(3,9), "leaf_size":[1,5], "weights":["uniform","distance"], "algorithm":["auto","ball_tree","brute"]	“algorithm”: [ball_tree], “leaf_size”: [1], “n_neighbors”: [8], “weights”: [distance]
AB	"n_estimators":[1,5,10,20,25,40,50,60], "learning_rate":[0.0001,0.001,0.1,1.5], "random_state":[42]	“learning_rate”: [1.5], “n_estimators”: [40], “random_state”: [42]
ET	"max_depth":[None], "max_features":[1,3,10], "random_state":[42] "min_samp_split":[2,3,10], "min_samp_leaf":[1,3,10], "n_estimators":[100,300], "criterion":["gini"], "random_state":[42]	“bootstrap”: [False], “criterion”: [gini], “max_depth”: [None], “max_features”: [3], “min_samples_leaf”:[3], “min_sampl_split”: [2], “n_estimators”: [300], “random_state”: [42]

8.4 Results and Discussion

This section shows the findings of the 10-fold cross validation after optimization, the results of each model based on the established performance metrics, the results of confusion metrics and the comparison of each prediction model before and after grid search based hyperparameter technique is applying.

The first experiment of this research work is comparing the results of each model of 10-fold cross validation score before and after parameter optimization technique is applying. After observing the results of ten-fold CV results of each model, the experiment is done with the best CV score results. In Table 8.1 above, the CV score before optimization is presented. And the Results of accuracy after optimization is presented in Table 8.3. The reason why we used this technique is; there is a difference between each model after applying the ten-fold cross validation.

TABLE 8.3: The 10-fold cross validation score of all models after applying Hyperparameter optimization

	1	2	3	4	5	6	7	8
MODEL	GB	SVM	RF	ET	LR	KNN	DT	AB
CV-Score	87.65	87.60	87.60	87.70	82.82	83.23	86.04	80.95

The results of each model after ten-fold cross validation technique is applying, there is a difference in their score. Among the classifiers, ET, GB, SVM, and RF, have the highest accuracy after tuning hyper-parameters accordingly. If we chose our models based on cross validation scores, we would not get the above results of our best model instead we would end up choosing the others. So it is recommended to select best models based on accuracy after tuning hyper-parameters though it is computationally intensive. So we have our best models with their best hyper-parameters that produces best accuracy out of a model. That means if we retrain the models using their best hyper-parameters, we will be able to get the very same score that we got after tuning hyper-parameters (see Table 8.5 for results).

On the other hand, we do feature importance for tree based models. The importance of the feature describes the statistical significance of each attributes in the data in terms of its model effect. The method of computing each feature's importance depends on the model. Figure 8.2, Figure 8.3, and Figure 8.4 shows the importance of features among established models.

In the feature importance experiment, GB and AB (in particular) give some features no importance (zero importance). On the other hand, RF, DT, and ET give all the features more or less importance but it doesn't give zero importance to any features. These are the tree based models that have 'feature importance's' method by default. LR, KNN and SVM don't have this method. In this problem, SVM uses "rbf" kernel (only possible for linear kernel to plot feature importance), so it's not possible to view feature importance given by SVM.

In the experiment part of this study, an extensive set of experiment are carried out in order to validate the soundness of the proposed method. We conduct the experiments in to two separate phases: the first phase of the experiment has done with the default parameters of the models. All models used in this study is tested their prediction power without applying grid search hyper-parameter optimization (which used the default parameter). The results of this experiment is presented in Table 8.4. We used Table 8.4 for further comparison with the result of the proposed method. The second phase of our experiments are done by our proposed methods. All the models are tested their prediction power by applying the grid search hyper-parameter optimization technique. The results of the proposed method is also presented in Table 8.5. Results are stated by the standard evaluation metrics.

Table 8.5 shows that the result of this GSHPO of the eight models used in this paper. As Table 8.5, in the testing phase, Gradient Boosting (GB) is fitted and executed the model with best parameters, after GSHPO applied (see Table 8.3) and found **87.67%**, **86.68%**, **83.14%**, **87.32%**, and **93.63%** of accuracy, precision, recall, F1, and AUC-ROC score respectively. And we measures the average of the absolute errors using mean-absolute-error (MAE) with **0.1233%**, evaluates how a model can predict a continuous value using root-mean-squared-error (RMSE) with **0.3512%**, and measure the degree that the predicted value and the actual value using root-squared (R^2) with **0.4134%**. In the second model Support Vector Machine (SVM), the best fit parameter result shows that **87.37%**, **86.22%**, **82.97%**, **87.07%**, and **89.19%** of accuracy, precision, recall, F1, and AUC-ROC score respectively. The MAE, RMSE, and R^2 are also scored **0.1258**, **0.3547**, and **0.4018** respectively.

In the third model of random forest (RF), the best parameter scored after GSHPO in the given evaluation metrics of accuracy, precision, recall, f1-score, and AUC-ROC is scored **86.72%**, **84.31%**, **83.98%**, **86.68%**, and **92.70%** respectively. And also, the MAE, RMSE, and R^2 are **0.0247**, **0.1573** and **0.8823** accordingly. In fourth

TABLE 8.4: Results of evaluation according to the standard metrics for each model before Grid Search technique is applying

Model	Accu	Prec	Rec	F1-score	AUC	MAE	RSME	R^2
Gradient Boosting **(GB)**	**0.858**	0.861	0.801	0.851	0.936	0.265	0.411	0.498
Support Vector Machine **(SVM)**	**0.845**	0.851	0.822	0.841	0.871	0.209	0.435	0.510
Random Forest **(RF)**	**0.834**	0.815	0.823	0.834	0.917	0.032	0.211	0.912
Extra Tree **(ET)**	**0.827**	0.818	0.823	0.826	0.894	0.035	0.217	0.981
Logistic Regression **(LR)**	**0.800**	0.800	0.743	0.800	0.805	0.189	0.437	0.255
K-nearest Neighbor **(KNN)**	**0.807**	0.745	0.730	0.806	0.867	0.201	0.378	0.488
Decision Tree **(DT)**	**0.817**	0.734	0.731	0.812	0.801	0.097	0.264	0.331
AdaBoost **(AB)**	**0.765**	0.725	0.701	0.734	0.826	0.217	0.501	0.251

TABLE 8.5: GSHPO results of evaluation according to the standard metrics for each model

Model	Fitting Time	Scoring Time	Accu	Prec	Rec	F1-score	AUC	MAE	RSME	R^2
GB	5.905	0.023	**0.876**	0.867	0.831	0.873	0.936	0.123	0.351	0.413
SVM	139.20	6.082	**0.874**	0.862	0.829	0.871	0.892	0.126	0.355	0.402
RF	1.602	0.205	**0.867**	0.843	0.839	0.867	0.927	0.025	0.157	0.882
ET	4.984	0.241	**0.862**	0.837	0.831	0.861	0.904	0.025	0.157	0.882
LR	0.896	0.012	**0.828**	0.812	0.761	0.820	0.825	0.172	0.415	0.183
KNN	0.712	2.458	**0.827**	0.799	0.778	0.824	0.871	0.119	0.345	0.435
DT	0.201	0.010	**0.825**	0.793	0.787	0.824	0.808	0.023	0.157	0.882
AB	1.822	0.071	**0.807**	0.789	0.726	0.795	0.833	0.193	0.439	0.080

model Extra Tree (ET), with the performance evaluation index of accuracy, precision, recall, F1, and AUC-ROC the result obtained by best parameter of GSHPO is **86.17%**, **83.74%**, **83.12%**, **86.11%**, and **90.43%** scored respectively and **0.0247**, **0.1573**, and **0.8823** is scored by MAE, RMSE, and R^2 respectively. The fifth model of this experiment Logistic regression (LR) also scored **82.79%**, **81.18%**, **76.12%**, **82.04%**, and **82.54%** respectively. This LR obtained the MAE, RMSE, and R^2 with **0.1718**, **0.4145**, and **0.1829** respectively.

The sixth model of this experiment is k-nearest Neighbor (KNN). In KNN model, the result scored by applying the GSHPO of best parameter the performance metrics of accuracy, precision, recall, f1-score, and AUC-ROC is **82.71%**, **79.97%**, **77.76%**, **82.37%**, and **87.10%** respectively. As the same parameter KNN also scored the MAE, RMSE, and R^2 is **0.1187**, **0.3446**, and **0.4352** respectively. At the seventh experiment the Decision Tree (DT) is tested. The DT model fitted and executed with best parameters like the above models with GSHPO. In the result obtained by accuracy, precision, recall, f1-score, and AUC-ROC metrics is **82.51**%, **79.32**%, **78.69**%, **82.41**%, and **80.80**% accordingly. And the other measure of this model, with MAE, RMSE, and R^2 is **0.0247**, **0.1573**, and **0.8823** respectively.

The last model tested in this experiment is AdaBoost (AB), in the AB the experiments are taken by GSHPO with best parameter and we found **80.70**%, **78.99**%, **72.64**%, **79.53**%, and **83.28**% of accuracy, precision, recall, F1, and AUC-ROC score respectively. And this model also recorded its prediction performance with **0.1934**, **0.4398**, and **0.0801** using MAE, RMSE, and R^2 respectively.

Table 8.5 also presents that the results of each model with it prediction performance and the experiments are taken with how each model executes their tasks with in the computational time. The computational time of the executions of the model in terms of fitting time and scoring time.

Table 8.4 and Table 8.5 presents the results of the experiment before GSHPO applying and the results of the proposed method respectively. Then, we see the effect of our methods to predict HIV test results by using the EDHS-HIV dataset. Our findings show that after tuning, the values of model hyper- parameters often differ from the default value. The tuning step often adjusts the value from the default to at least one hyper-parameter. As

we observe the results from the Table 8.4 and Table 8.5, there is a visible differences for each models before and after parameter optimization method is applying. All standard performance evaluation metrics are used to validate the experiments. Finally, we conclude that the Grid search hyper-parameter optimization is make a difference for predict the HIV test results.

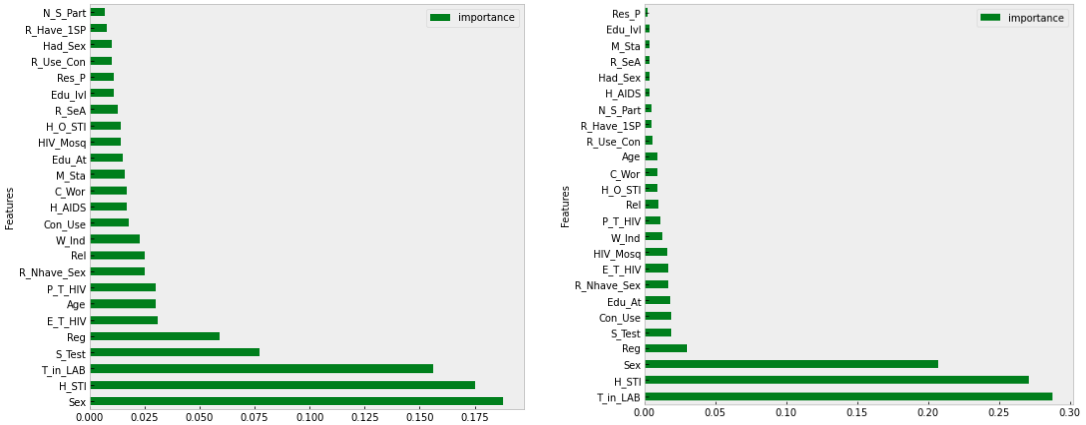

FIGURE 8.2: The feature importance of Random Forest and Decision Tree : RF (left), DT (right)

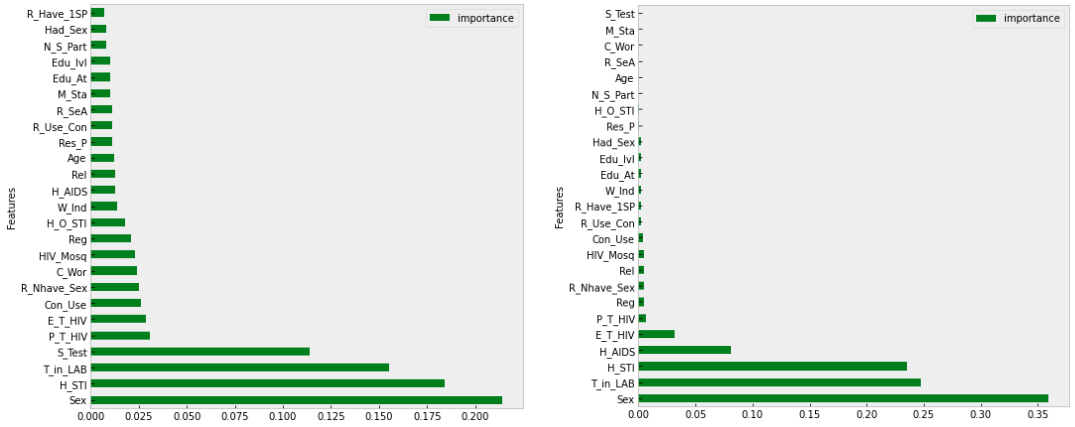

FIGURE 8.3: The feature importance of Extra Tree and Gradient Boosting : ET (left), GB (right)

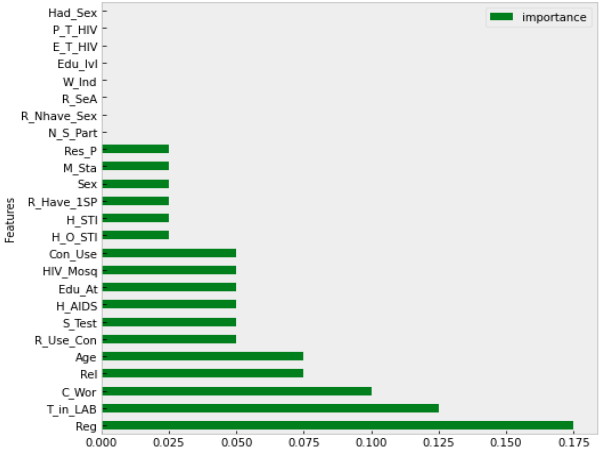

FIGURE 8.4: The feature importance of AdaBoost

We plot AUC-ROC to test the models. AUC-ROC is one of our performance metrics mentioned above in the proposed methodology section. It tells how much our model distinguishes between classes. Figure 8.5 shows

the AUC-ROC curve and the precision-recall tradeoff. The tradeoff that can be illustrated using precision-recall curve as a function of the decision threshold for the considered models.

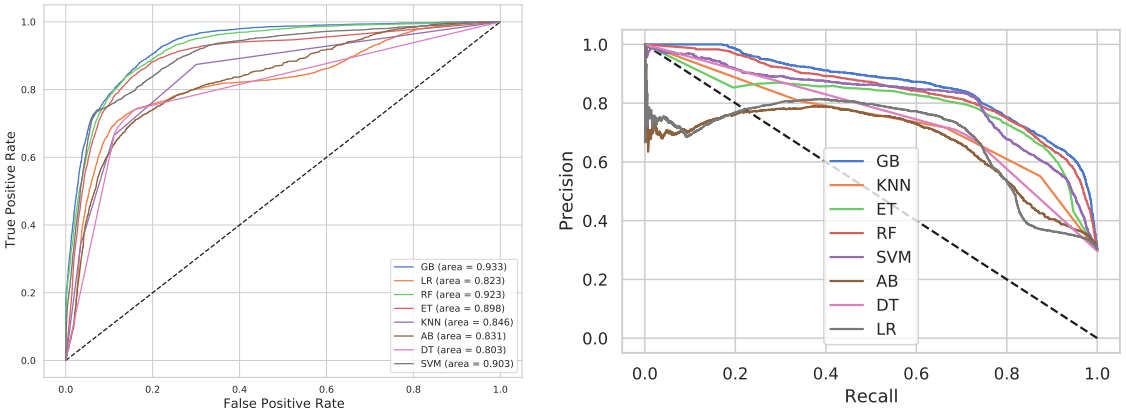

FIGURE 8.5: The comparison of each model with ROC Curve and Precision - Recall Curve: ROC (left), Precision - Recall (right)

The confusion matrix shows the number of correct classifications along with misclassifications when a model makes predictions for each class (positive or negative). We use the confusion matrix to validate our work, showing correctly and wrongly predicted samples in the test results. Figure 8.6 depicts how the confusion matrix is correctly predicted based on training data of the established model used in this paper. We focus to explain the three models of Confusion matrix results.

As Figure 8.6, For ET, the HIV test results correctly classified as Negative result (HIV–) are **69.48**% (true positives) and correctly classified as a positive result (HIV-+) are **28.04**% (true negatives). While **0.43**% (false positives) from class 0 (positive class) were misclassified as Negative and **2.04**% (false negatives) actually Negative were classified as Positive. And for GB, the HIV test results correctly classified as Negative (HIV–) are **66.01**% (true positives) and correctly classified as positive (HIV-+) is **21.65**% (true negatives). While **3.90**% (false positives) from class 0 (positive) were misclassified as negative and **8.44**% (false negatives) actually negative were misclassified as Positive. And for RF, the HIV test results correctly classified as Negative (HIV–) are **69.08**% (true positives) and correctly classified as positive (HIV-+) is 28.44% (true negatives). While **0.83**% (false positives) from class 0 (positive) were misclassified as negative and **1.65**% (false negatives) 0actually negative were misclassified as Positive. The experiments of this paper also do the learning curve of each model. To see their bias-variance trade-off of the model we plot the learning curve. Figure 8.7 depicts the learning curve of the eight models. As Figure 8.7, RF, DT, KNN, and ET are just doing okay. Among them, KNN is doing the best in terms of a bias-variance tradeoff since KNN's train accuracy and cross-validation accuracy are almost equal. Since training and validation, curves have not yet converged for these models, adding more instances might help. On the other hand, the learning curve of GB, AB, SVM, and LR indicates a little bit of high bias or low variance (underfitting) and as the curves have already converged, adding more training data just might not help. Rather adding more features and increasing the model's complexity might help.

8.5 Conclusion

In this paper, eight machine learning models were applied on the task of predicting HIV test results; we also evaluate which of the best parameter values can be used to obtain a relevant performance. To validate and compare the efficiency of our tuning process, we run an extensive set of experiments using the EDHS-HIV/AIDS

FIGURE 8.6: Confusion matrix results based on training data of the established mode

dataset with cross-validation. A comparison of the machine learning model was carried out to evaluate their results. We compare our proposed methods with the statistical way of optimization. This comparison is done through standard metrics such as accuracy, precision, recall, f1-score, AUC-ROC, MAE, RMSE, and R^2. Our experimental evaluation results of the model indicate that the GSHPO technique can be used to improve the prediction power and determine the problem of prediction. We observe that the tree-based models are performing better than the other models in these experiments. Our findings suggest that the hyper-parameters of the tuning model have a statistically important positive impact on the models' prediction accuracy. We also noticed that during the tuning process, most of the hyper-parameter values are modified and suggesting that default values are sub-optimal. Our results demonstrate that researchers in the prediction of HIV test results must take into account the optimization of hyper-parameters in their prediction, as theoretically it increases the prediction efficiency of machine learning models. The proposed method in this paper could be further applied for future research by using a larger collection of datasets and/or more costly algorithms to help the validation of the GSHPO under different conditions.

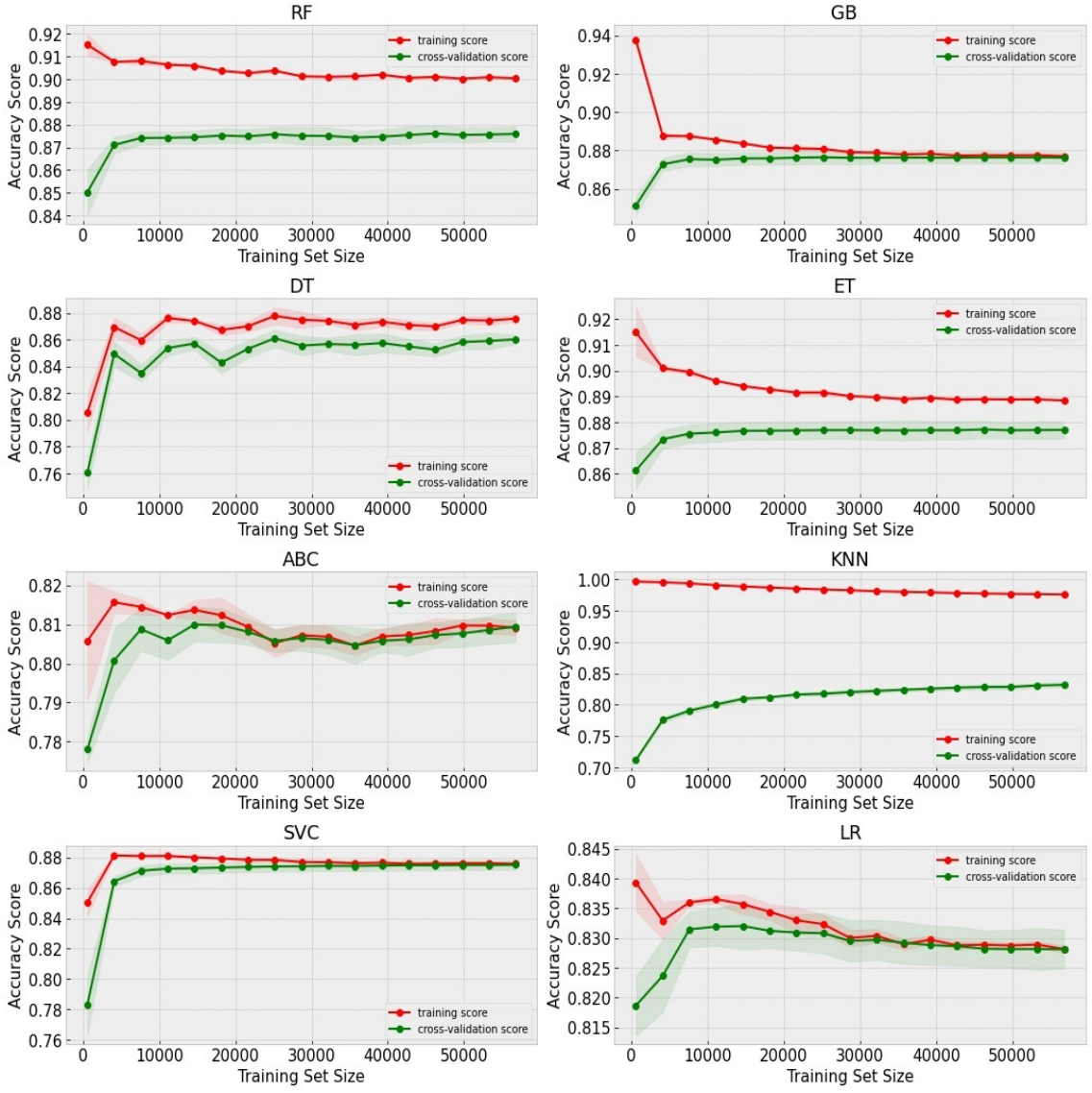

FIGURE 8.7: Bias-variance tradeoff of the model with learning curve

Chapter 9

Deep Learning Models for Predicting HIV Test Status using DHS data

9.1 Introduction

In this work, we design and deploy multiple deep learning application models to create a prediction system for HIV status. Six DL models have been created and are being used on Ethiopian demographic and health survey data (EDHS). We evaluated these deep learning models using well-known measures namely; accuracy, precision, recall, AUC, and F1-scores.

Our contributions are presented below in relation to the core goal of predicting HIV test status:

- Identification of the best performing deep learning algorithm for the task at hand among well-known and widely utilised ones;
  - To conduct an HIV test prediction research using deep learning application models with the EDHS dataset.
- We created the HIV/AIDS dataset and compared the findings using cutting-edge methodologies, and we got good findings.

9.2 Proposed Methodology

We discuss the proposed work in this part, which encompasses several phases such as pre-processing, normalizing features, and a deep learning-based prediction technique with parameter settings. Figure 9.1 depicts the architecture for the proposed deep learning models for predicting HIV test status in people using the EDHS dataset.

9.3 Experimental setup and result discussion

This part, describe the experimental setting, dataset description, performance measure and experimental findings and analysis.

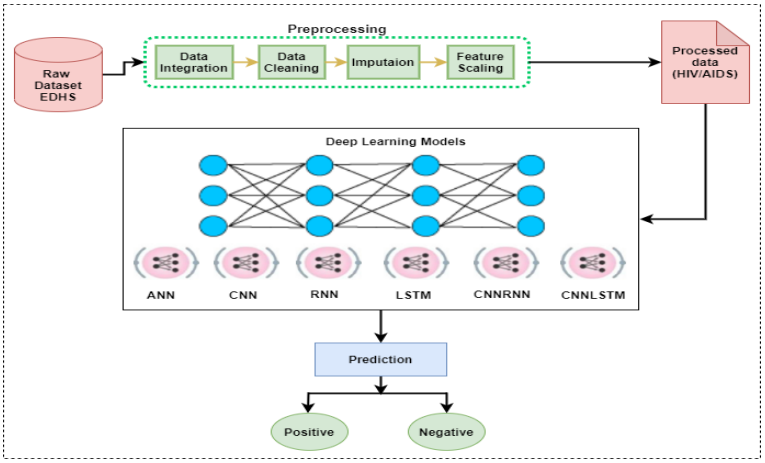

FIGURE 9.1: The proposed deep learning approach's architecture.

9.3.1 Experimental Setting

The parameters chosen and investigated for DL became: (1) **Default parameters**, those that remained constant throughout all tested configurations: (a) The Adam [57] optimization approach is used with momentum = 0.3 and 0.5, (b) the epochs is set to 100 for ANN and 250 for others, (c) mini-batch size of 256, (d) the Softmax is utilised as an output function, and (e) the decay was set to 0.00001. (2) **Variable parameters**, (a) ReLU as the activation function, (b) the binary entropy is a loss-function, (c) 0.001 as the learning rate, and (d) Dropout regularisation as the regularisation approach employed in this work. On each iteration, some of the neurons are dropped at random. This concept makes the network more versatile, removing some volatility and resulting in a more adaptive, less overfitted model. To establish the parameters for each DL model, we used a trial and error technique.

9.3.2 Experimental Result Analysis

With regard to the predictive performances, we discovered that best comprehensive recognised models on AUC score for predicting HIV test status were **89.72**% by ANN. It was discovered that predicting HIV test status from the EDHS dataset was considered as a difficult activity. Nonetheless, the best HIV test status prediction outcomes using ANN obtained reasonable accuracy of **85.5**%, precision of **84.4**%, recall of **85.7**%, and f1-score of **85.1**%.

TABLE 9.1: The outcomes of all DL models evaluated using a 10-fold cross-validation methodology.

Model	Acc	Pre	Rec	F1-score	AUC
ANN	**0.855**	**0.844**	**0.857**	**0.851**	**89.72**
CNN	0.834	0.833	0.845	0.841	87.52
RNN	0.847	0.821	0.822	0.834	87.62
LSTM	0.821	0.830	0.836	0.831	86.46
CNNLSTM	0.801	0.800	0.822	0.826	83.78
CNNRNN	0.792	0.800	0.821	0.801	83.77

In addition, as shown in Table 9.2 for the assessment results of DL models using a train-test split strategy, we examined the algorithm's performance using an 80:20 train-test split strategy. Even though the k-fold CV strategy is often employed within machine learning and deep learning research, particularly with small data. All algorithms outperformed with compared to the 10 fold CV technique on the subject of predictive performance,

with the best-performing algorithm, the RNN model, achieving an accuracy of **0.87**, precision of **0.87**, recall of **0.87**, f1-score of **0.87**, and AUC of **0.94**. As shown in Table 9.2, all DL models accuracy results were at least **0.834** or above. With **0.87**, the RNN model had the best evaluation performance. The CNNLSTM hybrid model was shown to be the second best model with **86.2**%.

TABLE 9.2: The evaluation outcomes of all DL models using the train-test split method.

Model	Training Acc	Testing Acc	Pre	Rec	F1-score	AUC	logloss error
ANN	0.881	0.861	0.858	0.858	0.858	0.94	0.3137
CNN	0.879	0.856	0.856	0.856	0.856	0.94	0.3201
RNN	**0.910**	**0.870**	**0.871**	**0.876**	**0.876**	**0.94**	**0.2909**
LSTM	0.870	0.844	0.844	0.844	0.844	0.92	0.3587
CNNLSTM	0.881	0.862	0.862	0.862	0.862	0.93	0.3130
CNNRNN	0.863	0.834	0.834	0.834	0.834	0.93	0.3707

Performance measure metrics values were found to be more than **83.0**%. Precision is the proportion of accurately predicted positive findings to the total number of expected positive findings. A perfect precision in information retrieval experiments should be 1. The greatest precision score in this study was obtained using RNN, which was **0.871**. A ratio of accurately predicted positive findings to all results is defined as recall. A recall score, like accuracy, must be one in order for the categorization process to be perfect. With **0.876**, the best recall value was attained using the RNN model. F1-score calculated as the weighted average of accuracy and recall scores. This criteria considers for both FP and FN. A high F1-score indicates that the predictor has few FP and few FN. In this scenario, the predictor identifies serious threats while avoiding false alarms. When the value of an F1-score is 1, it is deemed perfect. The best F1-score got with RNN was **0.876**, as with any other assessment criterion. In classification analysis, The AUC is used to determine best algorithms used to predicts target classes. In general, a score value of AUC **0.5** indicates that no variance, score in between **0.6** to **0.8** is held as an allowable, a score of **0.8** to **0.9** is regarded excellent, and a value of greater than 0.9 is regarded exceptional [58]. The AUC values of all DL models were outstanding since all of the outcomes were more than **0.9**. All DL models may be used to predict HIV test results based on their AUC values.

True positive rates are critical in health investigations since recall indicates the percentage of actual positives identified [59]. Recall is a significant assessment criterion in this study since it is computed by dividing number of properly recognised HIV positive samples by the total number of HIV test results. Besides, AUC score, plays an important role in health research since it has a relevant interpretation for health prediction [60]. Accuracy is a study criterion that indicates how near the sample parameters are to population characteristics. We can demonstrate that the study is generalizable, dependable, and valid by testing the correctness of the models [61]. As a result, just these three assessment indicators were examined in this study. The remaining ones were computed in order to compare the findings to earlier studies. The AUC values using the train test split strategy are shown in Figure 9.2, Figure 9.3, and Figure 9.4.

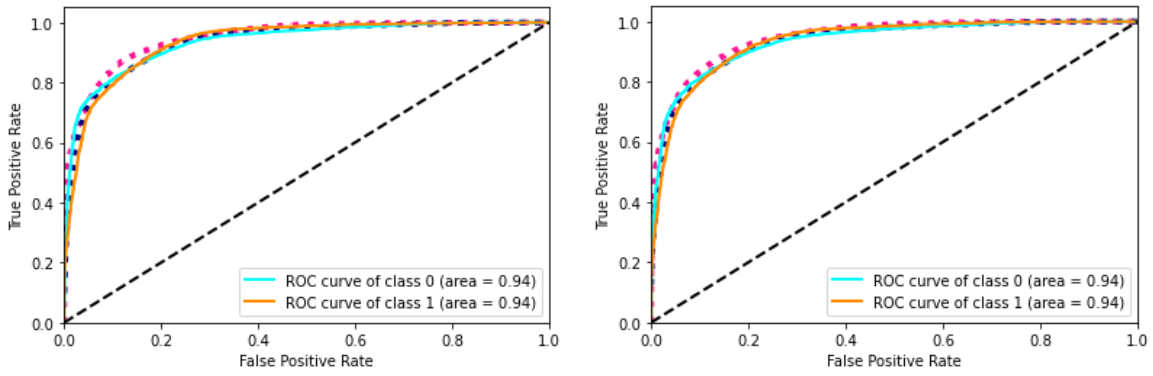
True Positive Rate
False Positive Rate
ROC curve of class 0 (area = 0.94)
ROC curve of class 1 (area = 0.94)
True Positive Rate
False Positive Rate
ROC curve of class 0 (area = 0.94)
ROC curve of class 1 (area = 0.94)

FIGURE 9.2: Deep learning Models AUC using the train-test split strategy: ANN (left), CNN (right)

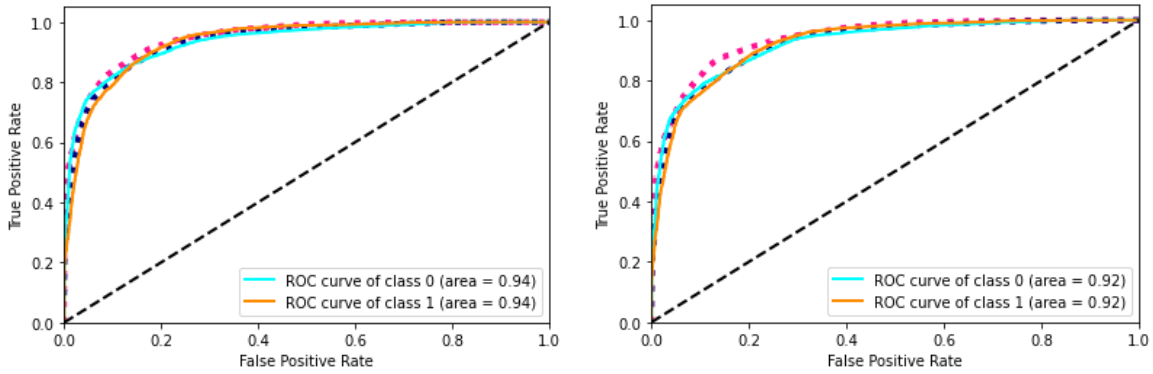
True Positive Rate
False Positive Rate
ROC curve of class 0 (area = 0.94)
ROC curve of class 1 (area = 0.94)
True Positive Rate
False Positive Rate
ROC curve of class 0 (area = 0.92)
ROC curve of class 1 (area = 0.92)

FIGURE 9.3: Deep learning Models AUC using the train-test split strategy: RNN (left), LSTM (right)

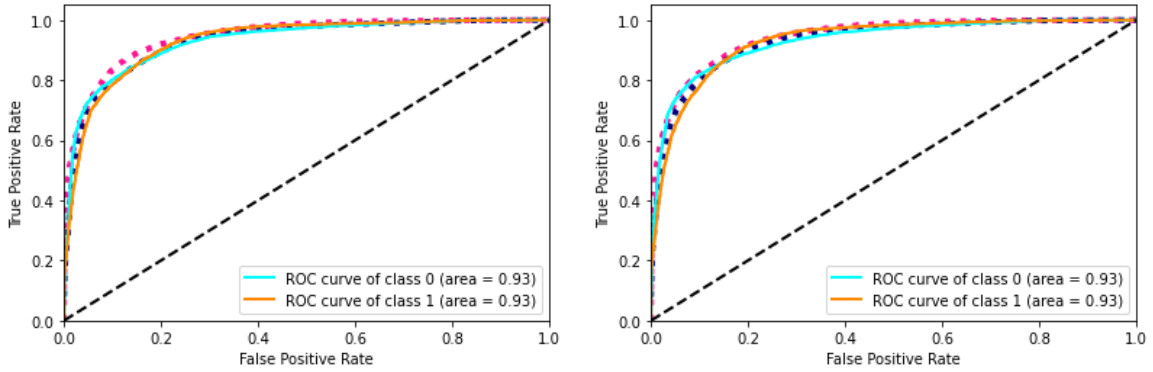
True Positive Rate
False Positive Rate
ROC curve of class 0 (area = 0.93)
ROC curve of class 1 (area = 0.93)
True Positive Rate
False Positive Rate
ROC curve of class 0 (area = 0.93)
ROC curve of class 1 (area = 0.93)

FIGURE 9.4: Deep learning Models AUC using the train-test split strategy: CNNLSTM (left), CNNRNN (right)

In addition to the metrics listed above, we calculated prediction accuracy to assess the efficacy of the proposed approach. Figure 9.5, Figure 9.6, and Figure 9.7 depicts the prediction accuracy of proposed technique on the HIV/AIDS dataset in terms of each DL model. Because of the flawless prediction accuracy on HIV/AIDS data, substantial difference is not there among lines related with training and test data, as shown in the Figure 9.5, Figure 9.6, and Figure 9.7.

We also used the log-loss error function to evaluate our work. As shown in Figure 9.8, Figure 9.9, and Figure 9.10, the training sample loss is near to 0, while the stated loss with the test sample is **0.3707** (refer Table 9.2), implying that more research on this specific dataset is required to reduce the error. As demonstrated in Loss Figures, the proposed technique outperforms all DL models by scoring the least number of errors in the test instances, with the ANN, CNN, RNN, LSTM, CNNLSTM, and CNNRNN scoring **0.3137**, **0.3201**, **0.2929**,

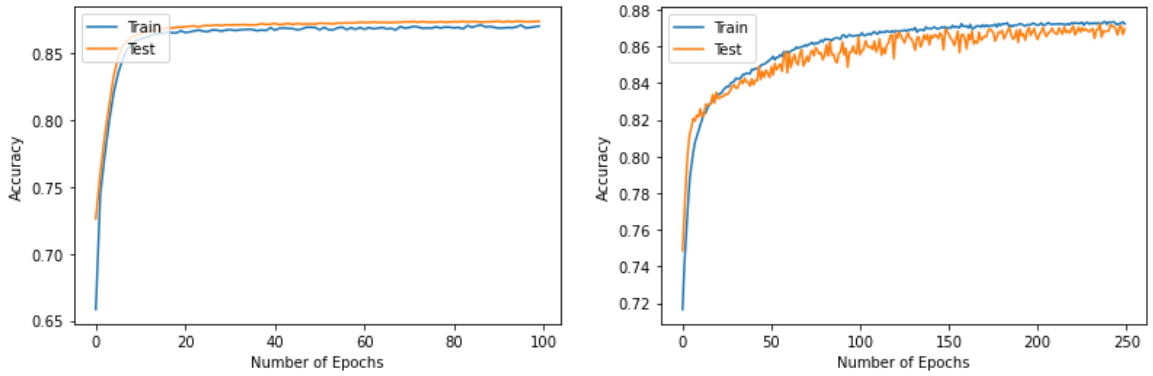

FIGURE 9.5: Prediction accuracy on the proposed deep learning model on HIV/AIDS datasets: ANN (left), CNN (right)

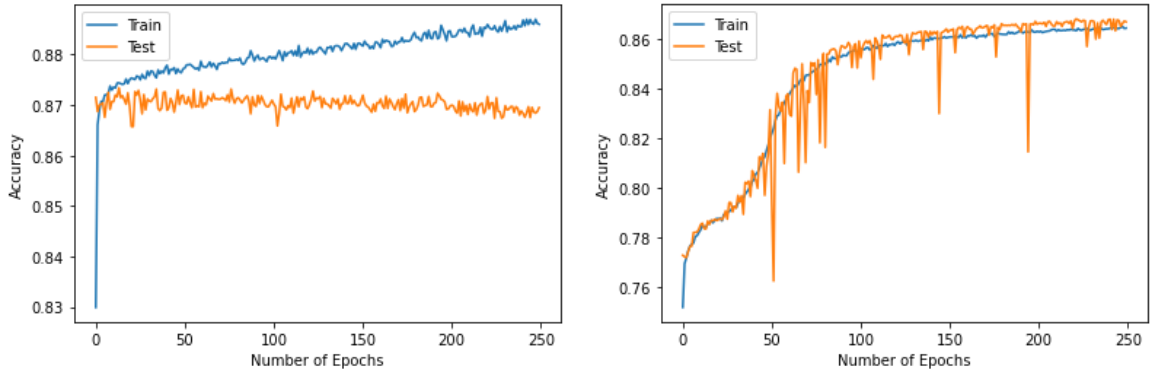

FIGURE 9.6: Prediction accuracy on the proposed deep learning model on HIV/AIDS datasets: RNN (left), LSTM (right)

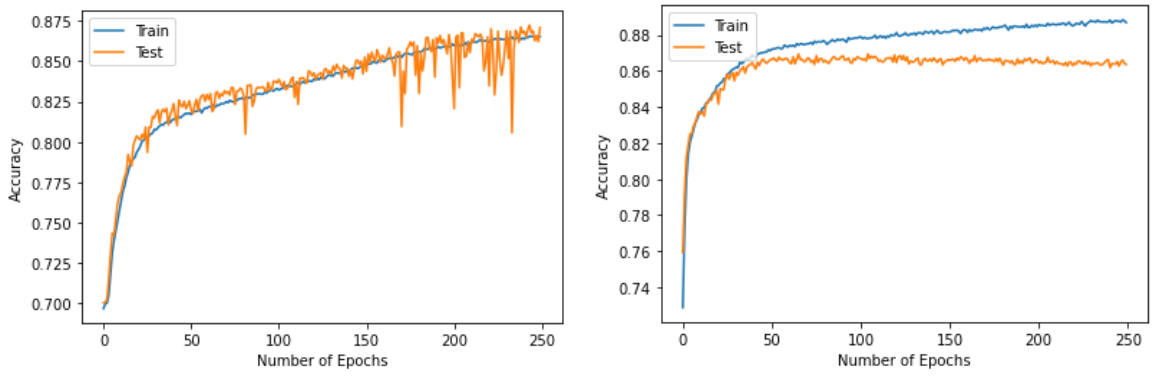

FIGURE 9.7: Prediction accuracy on the proposed deep learning model on HIV/AIDS datasets: CNNLSTM (left), CNNRNN (right)

0.3587, **0.3130**, and **0.3707**, correspondingly (refer Table 9.2). It is noticed that a distance between the training and the test lines indicates whether or not the model is over-fitting.

9.4 Comparison

This section of the study compares the proposed method to certain selected recent research in terms of performance measures. Table 9.3 compares the proposed method's assessment metrics to those of six other recent research works. The hyphen (-) in the table's specific cells indicates that the researchers did not consider metrics

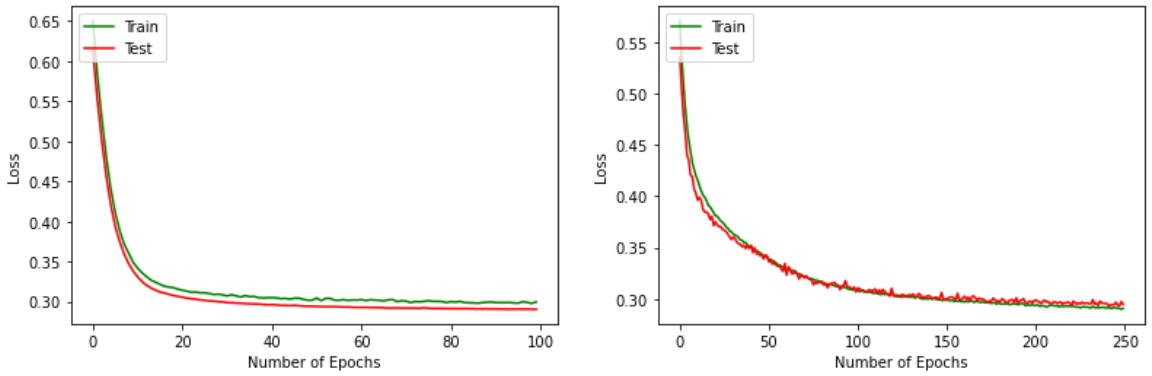

FIGURE 9.8: The loss caused by the proposed technique on deep learning models on HIV/AIDS datasets: ANN (left), CNN (right)

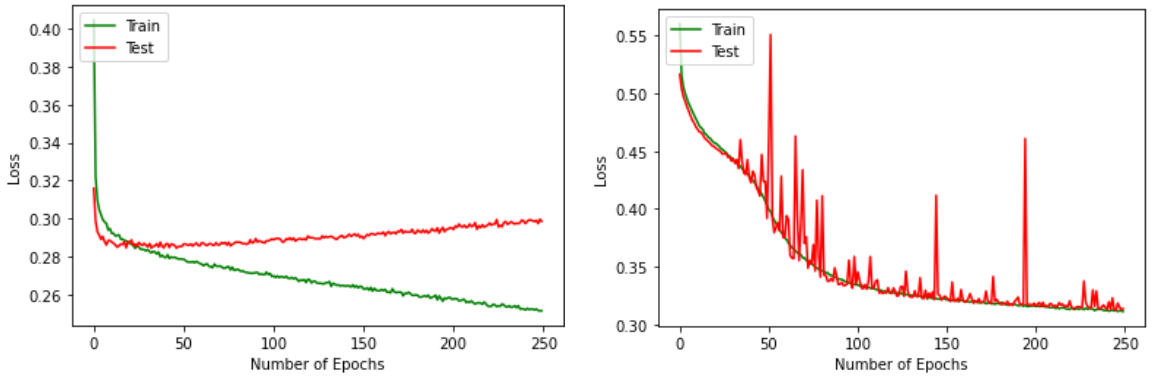

FIGURE 9.9: The loss caused by the proposed technique on deep learning models on HIV/AIDS datasets: RNN (left), LSTM (right)

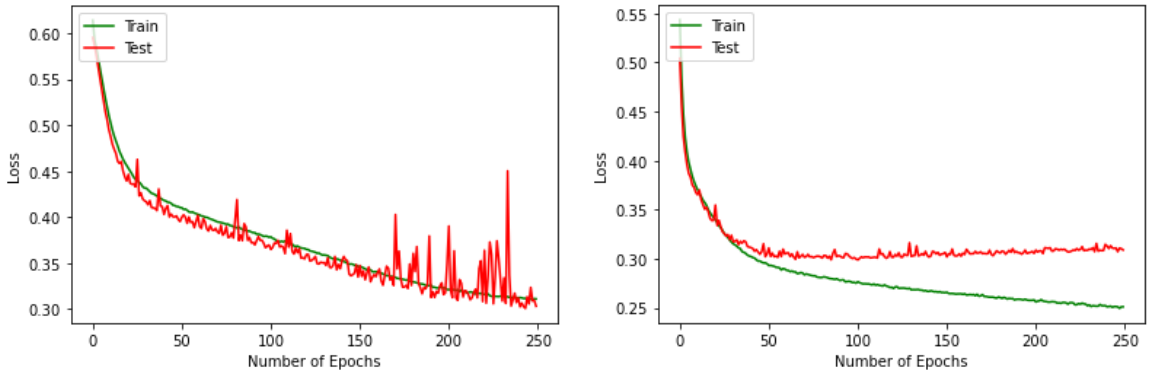

FIGURE 9.10: The loss caused by the proposed technique on deep learning models on HIV/AIDS datasets: CNNLSTM (left), CNNRNN (right)

in their study. As shown in Table 9.3, the best results were obtained with various models. Nonetheless, we has not employ ML in our research. We created six DL models and achieved higher accuracy, f1-scores, and AUC when compared to earlier similar efforts. In the considered HIV/AIDS dataset, the suggested technique obtains improved prediction performance, with **0.87** in total accuracy and f1-score and **0.94** in AUC score.

TABLE 9.3: The proposed model is compared to some of the most recent related research works.

References	Dataset	Methods	Best Result by	Accuracy	F1-score	AUC
McSharry, P. E., et al. [41]	Population-based HIV Impact Assessment	ML	XGBoost	-	0.789	-
Orel, E., et al. [42]	DHS data	ML	SVM	0.800	-	-
Ahlström, M. G., et al. [33]	Danish National Hospital Registry	ML	LR	-	-	88.4
Lu, X., et al. [43]	UCI ML repository	DL	RNN	-	**0.927**	-
Steiner, M. C., et al. [44]	HIV Drug Resistance database	DL	MLP	0.826	0.732	0.90
Betechuoh, B. L., et al. [62]	Antinatal Survey	DL	ANN	0.840	-	0.86
Proposed Method	**EDHS HIV/AIDS**	**DL**	**RNN**	**0.87**	**0.87**	**0.94**

9.5 Conclusion

In this work, deep learning models based on the EDHS dataset were used to predict HIV test results. Six deep learning models were used to analyse HIV/AIDS data. The data were normalised in the first stage of the study then utilised as an inputs for the DL models. Following that, prediction is performed, and the models' results were evaluated using precision, recall, accuracy, AUC, and F1- scores. We used 10 fold CV and train test split techniques to assess the models. In a 10-fold CV technique, the ANN deep learning model produced the most meaningful results, with an accuracy of **85.5**%, a recall of **85.7**%, and an AUC score of **87.72**%. Despite its popularity, this validation did not produce the best validation results. In the train-test split technique, the greatest accuracy, precision, recall, and AUC values were obtained with the RNN model, which were **87**%, **87**%, **87**%, and **94**%, respectively. Accuracy of all DL models produced in the study was greater than **83**%. Precision and recall values can be inferred in the same way. Finally, we discovered evidence that DL models may be used to predict HIV test status using demographic and health survey datasets. Our findings on the role of DHS in predicting HIV test status for people improve our knowledge of the consequences of HIV epidemics. Based on the findings of our study, we believe that health domain should investigate the use of DL models that analyse individual HIV test status in order to enhance and re-evaluate health policies and intervention mechanisms.

LIST OF PUBLICATIONS

Journal Publications

1. Daniel Mesafint and Manjaiah D H: **Paper Title**,– **"Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results"** *International Journal of Computers and Applications*, 2021, pp.:1-13,
doi:10.1080/1206212X.2021.1974663.
Available on: https://www.tandfonline.com/eprint/8UQNYCCVRNWIM6XCRHPM/full?target=10.1080/1206212X.2021.1974663.

2. Daniel Mesafint and Manjaiah D H: **Paper Title**,– **"Wrapper Based Feature Selection Techniques On EDHS-HIV/AIDS Dataset"** *European Journal of Molecular & Clinical Medicine (EJMCM)*, (ISSN:2515-8260), Volume-7 Issue-8, 2020 pp.2642-2657.
Available on: https://ejmcm.com/article_4776.html

3. Daniel Mesafint and Manjaiah D H: **Paper Title**,– **"Feature Selection Methods for Prediction of the Individual's Status of HIV/AIDS from EDHS dataset - A Filter Approach"** *International Journal of Research in Engineering & Advanced Technology (IJREAT)*, (ISSN:2320 – 8791), Volume-7 Issue-3, July - 2020 pp.29 - 36. Available on: http://www.ijreat.org/Issue39.html

4. Daniel Mesafint and Manjaiah D H: **Paper Title**,– **"Predictive Models of HIV/AIDS Epidemic Status using Data Mining Techniques: A Review"** *Journal of Emerging Technologies and Innovative Research (JETIR)*, (ISSN:2349-5162)(Online), Volume-6 Issue-5, May-2019 pp.10 - 16.
Available on: https://www.jetir.org/view?paper=JETIRCU06003

International Conferences

1. Daniel Mesafint Belete and Manjaiah D. Huchaiah, **Paper Title**,– **"Performance Evaluation of Classification Models for HIV/AIDS Dataset"**,– *Data Management, Analytics and Innovation. Lecture Notes on Data Engineering and Communications Technologies, 2021, vol 70. Springer, Singapore, doi:10.1007/978-981-16-2934-1_7) Springer*
Available on: https://link.springer.com/chapter/10.1007/978-981-16-2934-1_7

2. D. M. Belete and M. D H, **Paper Title**,– **"Classification of Binary Class HIV/AIDS Test Results Using Ensemble Learning Models"**,– *2020 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 2020, pp. 1240-1247, doi: 10.1109/DASA51403.2020.9317082.) IEEE*
Available on: https://ieeexplore.ieee.org/document/9317082

3. D. M. Belete and D. H. Manjaiah, **Paper Title**,– **"A Comparative Study of Filter and Wrapper Methods on EDHS – HIV/AIDS Dataset"**,– *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 1264-1271, doi: 10.1109/ICSSIT48917.2020.9214212.) IEEE.*
Available on: https://ieeexplore.ieee.org/abstract/document/9214212

Papers under Communication

1. **"A Deep Learning Algorithm for Modeling and Predicting of HIV Test Results using EDHS Dataset"**– *SN Computer Science* (**Under Communication**)

Bibliography

[1] WHO. *AIDS epidemic update, Joint United Nations Programme on HIV/AIDS*, 2006 (accessed December 13, 2018). URL https://www.hiv.gov/hiv-basics/overview/about-hiv-and-aids/what-are-hiv-and-aids.

[2] CSA. Ethiopia demographic and health survey 2016: Hiv report. addis ababa, ethiopia, and rockville, maryland, usa: Csa and icf, .

[3] CSA. Ethiopia demographic and health survey 2011: Hiv report. addis ababa, ethiopia, icf international calverton, maryland, usa, .

[4] Central Statistics Agency (CSA). Demographic and health survey, 2018 (accessed on October 28, 2018). URL http://www.csa.gov.et/survey-report/category/2-demographic-and-health-survey.

[5] Pinar Yildirim. Filter based feature selection methods for prediction of risks in hepatitis disease. *International Journal of Machine Learning and Computing*, 5(4):258, 2015.

[6] Asha Gowda Karegowda, MA Jayaram, and AS Manjunath. Feature subset selection using cascaded ga & cfs: A filter approach in supervised learning. *International Journal of Computer Applications*, 975:8887, 2011.

[7] Subhajit Dey Sarkar and Saptarsi Goswami. Empirical study on filter based feature selection methods for text classification. *International Journal of Computer Applications*, 81(6), 2013.

[8] R Nair and A Bhagat. Feature selection method to improve the accuracy of classification algorithm. *Int. J. Innov. Technol. Explor. Eng*, 8(6):124–127, 2019.

[9] Hany M Harb and Abeer S Desuky. Feature selection on classification of medical datasets based on particle swarm optimization. *International Journal of Computer Applications*, 104(5), 2014.

[10] Mehnaz Khan and SMK Quadri. Effects of using filter based feature selection on the performance of machine learners using different datasets. *BVICA M's International Journal of Information Technology*, 5 (2):597, 2013.

[11] D Chitra and GM Nasira. Wrapper based feature selection for ct image. *ICTACT journals in/paper/IJIVP_Paper_7*, pages 1096–1103, 2015.

[12] Rattanawadee Panthong and Anongnart Srivihok. Wrapper feature subset selection for dimension reduction based on ensemble learning algorithm. *Procedia Computer Science*, 72:162–169, 2015.

[13] Asha Gowda Karegowda, MA Jayaram, and AS Manjunath. Feature subset selection problem using wrapper approach in supervised learning. *International journal of Computer applications*, 1(7):13–17, 2010.

[14] Akin Ozcift and Arif Gulten. A robust multi-class feature selection strategy based on rotation forest ensemble algorithm for diagnosis of erythemato-squamous diseases. *Journal of medical systems*, 36(2):941–949, 2012.

[15] MJ Abinash and V Vasudevan. A study on wrapper-based feature selection algorithm for leukemia dataset. In *Intelligent Engineering Informatics*, pages 311–321. Springer, 2018.

[16] Luis Talavera. An evaluation of filter and wrapper methods for feature selection in categorical clustering. In *International Symposium on Intelligent Data Analysis*, pages 440–451. Springer, 2005.

[17] N Nnamoko, F Arshad, D England, J Vora, and J Norman. Evaluation of filter and wrapper methods for feature selection in supervised machine learning. *Age*, 21(81):33–2, 2014.

[18] José M Cadenas, M Carmen Garrido, and Raquel MartíNez. Feature subset selection filter–wrapper based on low quality data. *Expert systems with applications*, 40(16):6241–6252, 2013.

[19] Hu Min and Wu Fangfang. Filter-wrapper hybrid method on feature selection. In *2010 Second WRI Global Congress on Intelligent Systems*, volume 3, pages 98–101. IEEE, 2010.

[20] Pengyi Yang, Wei Liu, Bing B Zhou, Sanjay Chawla, and Albert Y Zomaya. Ensemble-based wrapper methods for feature selection and class imbalance learning. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 544–555. Springer, 2013.

[21] Meherwar Fatima, Maruf Pasha, et al. Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, 9(01):1, 2017.

[22] Peter A Idowu, Theophlus A Aladekomo, Olutola Agbelusi, Olumuyiwa B Alaba, and Jeremiah A Balogun. Prediction of pediatric hiv/aids survival in nigeria using naïve bayes' approach. *International Journal of Child Health and Human Development*, 10(2):131, 2017.

[23] Yunuscan Koçak, Tansel Özyer, and Reda Alhajj. Classification of hiv data by constructing a social network with frequent itemsets. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 949–953. IEEE, 2016.

[24] C Olayemi Olufunke, O Olasehinde Olayemi, and O Agbelusi. Predictive model of pediatric hiv/aids survival in nigeria using support vector machine. *Communications*, 5:29–36.

[25] Alexandra Oliveira, Brígida Mónica Faria, A Rita Gaio, and Luís Paulo Reis. Data mining in hiv-aids surveillance system. *Journal of medical systems*, 41(4):51, 2017.

[26] Hai-lei WU, Ji-sheng QIAN, Chun ZHANG, et al. A hiv carrier forecasting model for quarantine based on support vector machines [j]. *Practical Preventive Medicine*, 11, 2010.

[27] Brain Leke-Betechuoh, Tshilidzi Marwala, Taryn Tim, and Monica Lagazio. Prediction of hiv status from demographic data using neural networks. In *2006 IEEE International Conference on Systems, Man and Cybernetics*, volume 3, pages 2339–2344. IEEE, 2006.

[28] Patrick Igulot and Monica A Magadi. Socioeconomic status and vulnerability to hiv infection in uganda: Evidence from multilevel modelling of aids indicator survey data. *AIDS research and treatment*, 2018, 2018.

[29] CD Anisha and N Arulanand. Early prediction of parkinson's disease (pd) using ensemble classifiers. In *2020 International Conference on Innovative Trends in Information Technology (ICITIIT)*, pages 1–6. IEEE, 2020.

[30] BH Shekar and Guesh Dagnew. Grid search-based hyperparameter tuning and classification of microarray cancer data. In *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, pages 1–8. IEEE, 2019.

[31] Rafael G Mantovani, Tomáš Horváth, Ricardo Cerri, Joaquin Vanschoren, and André CPLF de Carvalho. Hyper-parameter tuning of a decision tree induction algorithm. In *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 37–42. IEEE, 2016.

[32] Philipp Probst, Bernd Bischl, and Anne-Laure Boulesteix. Tunability: Importance of hyperparameters of machine learning algorithms. *arXiv preprint arXiv:1802.09596*, 2018.

[33] Magnus G Ahlström, Andreas Ronit, Lars Haukali Omland, Søren Vedel, and Niels Obel. Algorithmic prediction of hiv status using nation-wide electronic registry data. *EClinicalMedicine*, 17:100203, 2019.

[34] Dennis Adjei Annang. *Performance Comparison of Data Mining Techniques for Predicting Hiv Status Among Female Sex Workers in Ghana*. PhD thesis, University Of Ghana, 2018.

[35] Tesfay Gidey Hailu et al. Comparing data mining techniques in hiv testing prediction. *Intelligent Information Management*, 7(03):153, 2015.

[36] Girma Awoke. Predicting hiv infection risk factor using volantory counseling and testing data: A case of african aids initiative international, 2012. unpublished thesis.

[37] Teklu Urgesa. Application of data mining technology on antiretroviral therapy(art) data: a case of adama and assella hospital, 2010. unpublished thesis.

[38] Birhanu Fetene. Application of data mining for effective prognosis of hiv/aids testing, 2016. unpublished thesis.

[39] Elias Lemuye. *HIV status predictive modeling using data mining technology*. Lap Lambert Academic Publ, 2012.

[40] Yining Bao, Nicholas A Medland, Christopher K Fairley, Jinrong Wu, Xianwen Shang, Eric PF Chow, Xianglong Xu, Zongyuan Ge, Xun Zhuang, and Lei Zhang. Predicting the diagnosis of hiv and sexually transmitted infections among men who have sex with men using machine learning approaches. *Journal of Infection*, 82(1):48–59, 2021.

[41] Patrick E McSharry, Charles Mutai, Innocent Ngaruye, and Edouard Musabanganji. Use of machine learning techniques to identify hiv predictors for screening. 2020.

[42] Erol Orel, Rachel Esra, Janne Estill, Stephane Marchand-Maillet, Aziza Merzouki, and Olivia Keiser. Machine learning to identify socio-behavioural predictors of hiv positivity in east and southern africa. *medRxiv*, 2020.

[43] Xinyu Lu, Lifang Wang, and Zejun Jiang. The application of deep learning in the prediction of hiv-1 protease cleavage site. In *2018 5th International Conference on Systems and Informatics (ICSAI)*, pages 1299–1304. IEEE, 2018.

[44] Margaret C Steiner, Keylie M Gibson, and Keith A Crandall. Drug resistance prediction using deep learning techniques on hiv-1 sequence data. *Viruses*, 12(5):560, 2020.

[45] G Wang, W Wei, J Jiang, C Ning, H Chen, J Huang, B Liang, N Zang, Y Liao, R Chen, et al. Application of a long short-term memory neural network: a burgeoning method of deep learning in forecasting hiv incidence in guangxi, china. *Epidemiology & Infection*, 147, 2019.

[46] Birru Asmare. Application of data mining technology to support vct for hiv: A case of center for disease control and prevention, 2009. unpublished thesis.

[47] AnHai Doan, Alon Halevy, and Zachary Ives. *Principles of data integration*. Elsevier, 2012.

[48] Luai Al Shalabi, Zyad Shaaban, and Basel Kasasbeh. Data mining: A preprocessing engine. *Journal of Computer Science*, 2(9):735–739, 2006.

[49] Slobodan Ribaric and Ivan Fratric. Experimental evaluation of matching-score normalization techniques on different multimodal biometric systems. In *MELECON 2006-2006 IEEE Mediterranean Electrotechnical Conference*, pages 498–501. IEEE, 2006.

[50] George Forman et al. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, 3(Mar):1289–1305, 2003.

[51] Naoual El Aboudi and Laila Benhlima. Review on wrapper feature selection approaches. In *2016 International Conference on Engineering & MIS (ICEMIS)*, pages 1–5. IEEE, 2016.

[52] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[53] Chris Thornton, Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 847–855, 2013.

[54] Daniel Mesafint Belete and DH Manjaiah. A comparative study of filter and wrapper methods on edhs–hiv/aids dataset. In *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 1264–1271. IEEE, 2020.

[55] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[56] MohammadNoor Injadat, Abdallah Moubayed, Ali Bou Nassif, and Abdallah Shami. Systematic ensemble model selection approach for educational data mining. *Knowledge-Based Systems*, 200:105992, 2020.

[57] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[58] Jayawant N Mandrekar. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316, 2010.

[59] Anand Avati, Kenneth Jung, Stephanie Harman, Lance Downing, Andrew Ng, and Nigam H Shah. Improving palliative care with deep learning. *BMC medical informatics and decision making*, 18(4):55–64, 2018.

[60] Adina Najwa Kamarudin, Trevor Cox, and Ruwanthi Kolamunnage-Dona. Time-dependent roc curve analysis in medical research: current methods and applications. *BMC medical research methodology*, 17 (1):1–19, 2017.

[61] Roger Pierce. Evaluating information: Validity, reliability, accuracy, triangulation. *Research Methods in Politics: A Practical Guide; Sage Publications: Edmonton, AB, Canada*, 2008.

[62] Brain Leke Betechuoh, Tshilidzi Marwala, and Thando Tettey. Autoencoder networks for hiv classification. *Current Science*, pages 1467–1473, 2006.

Research Scholar:

Daniel Mesafint Belete,
Research Scholar,
Department of Computer Science,
Mangalore University.

Research Supervisor:

Dr. Manjaiah D H,
Professor,
Department of Computer Science,
Mangalore University.